Universal Filtering via Hidden Markov Modeling *

Taesup Moon Tsachy Weissman

February 1, 2008

Abstract

The problem of discrete universal filtering, in which the components of a discrete signal emitted by an unknown source and corrupted by a known DMC are to be causally estimated, is considered. A family of filters are derived, and are shown to be universally asymptotically optimal in the sense of achieving the optimum filtering performance when the clean signal is stationary, ergodic, and satisfies an additional mild positivity condition. Our schemes are comprised of approximating the noisy signal using a hidden Markov process (HMP) via maximum-likelihood (ML) estimation, followed by the use of the forward recursions for HMP state estimation. It is shown that as the data length increases, and as the number of states in the HMP approximation increases, our family of filters attain the performance of the optimal distribution-dependent filter.

Index Terms - Universal filtering, finite alphabet, hidden Markov process (HMP), stochastic setting, randomized scheme, forward-backward recursion state estimation, ML parameter estimation

1 Introduction

The problem of estimating a discrete-time, finite-alphabet source signal $\{X_t\}_{t\in T}$ from the entire observation of a noisy signal $\{Z_t\}_{t\in T}$, which has been corrupted by a known discrete memoryless channel (DMC), has been thoroughly studied recently in [21]. It has been shown that even though the source distribution is unknown, an algorithm called DUDE can universally achieve the asymptotically optimal performance. This result has been extended in various directions such as the case of channel uncertainty [9], the case where the channel has memory [22], the case of non-discrete noisy signal components [6], and the case where the reconstruction is required to depend causally on the noisy signal [18][19]. In this paper, we revisit the last case, taking a different approach from [18][19].

The case where we estimate X_t causally based on observation of the noisy signal $Z^t = (Z_1, \cdots, Z_t)$, is referred to as *filtering*. The filter can be either deterministic or randomized (a concept that will be explained in detail later). In this paper, we will only focus on the *stochastic setting*, where we assume $\{X_t\}$ is a stationary and ergodic stochastic process. With the stochastic setting assumption, and under the same performance criterion of [21], i.e., minimizing the expected normalized cumulative loss, knowledge of the conditional distribution of X_t given Z^t at each time t is required to achieve the optimal performance. Also, by the same argument as in [21, Section III], this conditional distribution can be obtained by the conditional distribution of Z_t given Z^{t-1} when the invertible DMC is known. (We call a channel is invertible if its transition probability matrix is invertible.)

However, for the *universal* filtering setting, where the probability distribution of the source is unknown, the conditional distribution of Z_t given Z^{t-1} is also not known and need be learned from the observed noisy signal. Therefore, if we can learn this conditional distribution accurately as the observation length increases, we can hope to build the universal filtering scheme that achieves the asymptotically optimal performance from the estimated

*Authors are with the department of electrical engineering, Stanford University, Stanford, CA 94305, USA. e-mails: {tsmoon, tsachy}@stanford.edu. The material in this paper was partially presented at the 2005 ISIT, [17]. The research was partially supported by NSF grant CCR-0311633

conditional distribution. To pursue this goal, [18][19] adopt the universal prediction[15] approach. That is, they first get an estimate of the conditional distribution of Z_t given Z^{t-1} by employing a universal predictor for the observed noisy signal, and then by inverting the known DMC, obtain an estimate of the conditional distribution of X_t given Z^t.

Unlike the approach of [18][19], in this work, we turn our attention to the rich theory of hidden Markov process (HMP) models to directly obtain a different kind of estimate of the conditional distribution of X_t given Z^t, without going through the channel inversion stage.

Generally, HMPs are defined as a family of stochastic processes that are outputs of a memoryless channel whose inputs are finite state Markov chains. As can be seen in [7], these HMP models arise in many areas, such as information theory, communications, statistics, learning, and speech recognition. Among these applications of HMPs, there are many situations where the state of the underlying Markov chain need be estimated based on the observed hidden Markov process. If the exact parameters of the HMP, namely, the state transition probability of the Markov chain and the channel transition density, as well as the order of the Markov chain are known, then this problem can be easily solved via well-known forward-backward recursions which were discovered by [4] and [2]. Especially, when we are estimating the state based on the causal observation of the HMP, we only need the forward recursion formula. In addition, much work has been done for the state estimation, where the order is known, but the parameters of the HMP are unknown. In this case, the parameters are first estimated via maximum likelihood (ML) estimation or the EM algorithm, then the state is estimated by using the estimated parameters in the recursion formula. A detailed explanation of this approach and the property of the ML parameter estimation can be found in [2][3][12][8]. Furthermore, this was extended to the case where the order of the Markov chain is also not known, but the upper bound on the order is known. In this case, the order estimation is first performed before the parameter and state estimation, and the above process is repeated. The references for the order estimation are given in [11][13][20]. There also has been work for the case where even the knowledge of the upper bound on the order of the Markov chain is not required[8][23].

From these rich theories for the state, parameter, and order estimation of HMPs, we can see that it is possible to build a universal filtering scheme if the source distribution is known to be a finite state Markov chain. That is, since the channel is memoryless and fixed in our setting, if our source $\{X_t\}$ is a finite state Markov chain, then obviously, $\{Z_t\}$ is an HMP, and we can first estimate the order of the Markov chain, then estimate the parameter, and finally perform forward recursion to learn the conditional distribution of X_t given Z^t. From the consistency results of order estimation and parameter estimation, this conditional distribution will be an accurate estimate of the true one, and we can use it to build the universal filtering scheme.

Now, in our work, we extend this approach to the case where our source $\{X_t\}$ is a general stationary and ergodic process (with some benign conditions), which need not be a Markov source at all, and show that we can still build a universal filtering scheme that achieves asymptotically optimal performance. The skeleton of our scheme is the following: We first "model" our source as a finite state Markov chain with a certain order, or equivalently, model the noisy observed signal $\{Z_t\}$ as an HMP in a certain class. Then, we estimate the parameters of the HMP that "approximates" the noisy signal best in that class. We will show that from the consistency result about the ML parameter estimation for the mismatched model [8], these estimated parameters will give an accurate estimation of the conditional distribution of X_t given Z^t, as the observation length increases and the HMP class gets richer. Then, this result will guarantee that our universal filter using this conditional distribution will attain the asymptotically optimal performance. In practice, this approach has been heuristically employed in many applications for nonlinear filtering without theoretical justification. Therefore, this work shows the first theoretical proof on the justification

of the HMP-based universal filtering scheme.

The remainder of the paper is organized as follows. Section 2 introduces some notation and preliminaries that are needed for setting up the problem. In Section 3, the universal filtering problem is defined explicitly. In Section 4, our universal filtering scheme is devised, the main theorem is stated, and proved. Section 5 extends our approach to the case where the channel has memory. Section 6 concludes the paper and lists some related future directions. Detailed technical proofs that are needed in the course of proving our main results are given in the Appendix.

2 Notation and preliminaries

2-A General notation

We assume that the clean, noisy and reconstruction signal components take their values in the same finite M-ary alphabet $\mathcal{A} = \{0, \cdots, M-1\}$. The simplex of M-dimensional column probability vectors will be denoted as $\mathcal{M}$.

The DMC is known to the filter and is denoted by its transition probability matrix $\mathbf{\Pi} = \{\Pi(i,j)\}_{i,j\in\mathcal{A}}$. Here, $\Pi(i,j)$ denotes the probability of channel output symbol j when the input is i. We assume $\Pi(i,j) > 0\ \forall i,j$, and let $\Pi_{\min} = \min_{i,j}\Pi(i,j)$. We assume this channel matrix is invertible and denote the inverse as $\mathbf{\Pi}^{-1}$. Let Π_i^{-1} denote the i-th column of $\mathbf{\Pi}^{-1}$. We also assume a given loss function (fidelity criterion) $\Lambda : \mathcal{A}^2 \to [0,\infty)$, represented by the loss matrix $\mathbf{\Lambda} = \{\Lambda(i,j)\}_{i,j\in\mathcal{A}}$, where $\Lambda(i,j)$ denotes the loss incurred when estimating the symbol i with the symbol j. The maximum single-letter loss will be denoted by $\Lambda_{\max} = \max_{i,j\in\mathcal{A}}\Lambda(i,j)$, and λ_j will denote the j-th column of $\mathbf{\Lambda}$.

As in [21], we define the extended Bayes response associated with the loss matrix $\mathbf{\Lambda}$ to any column vector $\mathbf{V} \in \mathbb{R}^M$ as

$$B(\mathbf{V}) = \arg\min_{\hat{x}\in\mathcal{A}} \lambda_{\hat{x}}^T \mathbf{V},$$

where $\arg\min_{\hat{x}\in\mathcal{A}}$ denotes the minimizing argument, resolving ties by taking the letter in the alphabet with the lowest index.

We let P denote the true joint probability law of the clean and noisy signal, and $E(\cdot)$ denote expectation with respect to P. Also, every almost sure convergence is with respect to P. If we need to refer to the probability law of clean or noisy signal induced by P, we denote P_X and P_Z, respectively. If P is written in a bold face, $\mathbf{P}$, with a subscript, it stands for a simplex vector in $\mathcal{M}$ for the corresponding distribution of the subscript. For example, $\mathbf{P}_{X_t|z^t}$ is a column M-vector whose i-th component is $P(X_t = i|Z^t = z^t)$.

When we have some other probability law denoted as Q, and want to measure its difference from P, a natural choice of such a measure is the relative entropy rate. First, denote the n-th order relative entropy between P and Q as

$$D_n(P||Q) = \sum_{z^n} P(z^n)\log\frac{P(z^n)}{Q(z^n)} = E\Big(\log\frac{P(Z^n)}{Q(Z^n)}\Big).$$

Then, the relative entropy rate (also known as Kullback-Leibler divergence rate) is defined as

$$\mathbf{D}(P\|Q) \triangleq \lim_{n\to\infty}\frac{1}{n}D_n(P||Q)$$

if the limit exists. When Q is a probability law in a certain class of HMPs, this limit always exists and the relative entropy rate is well defined. A more detailed discussion about this limit will be given in Lemma 2. This relative entropy rate will play a central role in analyzing our universal filtering scheme.

2-B Hidden Markov Processes (HMPs)

2-B.1 Definition

As stated in the Introduction, the HMPs are generally defined as a family of stochastic processes that are outputs of a memoryless channel whose inputs are finite state Markov chains. Throughout the paper, we will only consider the case in which the alphabet of HMP, $\mathcal{Z}$, and underlying Markov chain, $\mathcal{X}$, are finite and equal, i.e., $\mathcal{Z} = \mathcal{X} = \mathcal{A}$, and the channel is DMC and invertible.

There are three parameters that determine the probability laws of HMP: π, the initial distribution of finite state Markov chain; A, the probability transition matrix of finite state Markov chain, and B, the probability transition matrix of DMC. The triplet $\{\pi, A, B\}$ is referred to as the parameter of HMP. Let Θ be a set of all θ's where $\theta := \{\pi_\theta, A_\theta, B_\theta\}$. For each θ, we can calculate the likelihood function

$$Q_\theta(z^n) = \pi_\theta \prod_{t=1}^{n} (\hat{\mathbf{B}}_{\theta,t} A_\theta)\mathbf{1},$$

where $\hat{\mathbf{B}}_{\theta,t}$ is $M \times M$ diagonal matrix whose (j,j)-th entry is the (j, z_t)-th entry of B_θ, and $\mathbf{1}$ is the $M \times 1$ vector with all entries equal to 1.

Now, let $\Theta_k \subset \Theta$ be a set of θ's, such that the order of underlying Markov chain of HMP is k. Furthermore, for some $\delta > 0$, define $\Theta_k^\delta \subset \Theta_k$ as the set of $\theta \in \Theta_k$ satisfying:

- $a_{ij,\theta} \geq \delta$, if the first $k-1$ components of the k-tuple state j are equal to the last $k-1$ components of k-tuple state i
- $a_{ij,\theta} = 0$, otherwise
- $b_{ij,\theta} = \mathbf{\Pi}(i,j)$, for $\forall i, j$,

where $a_{ij,\theta}$ is (i,j)-th entry of A_θ, and $b_{ij,\theta}$ is (i,j)-th entry of B_θ. In particular, if $\theta \in \Theta_k^\delta$ then: 1) the stochastic matrix A_θ is irreducible and aperiodic; thus, if the Markov chain is stationary, π_θ is the stationary distribution of the Markov chain, and is uniquely determined from A_θ, 2) $B_\theta = \mathbf{\Pi}$ $\forall \theta$, and, therefore, θ is completely specified by A_θ. For notational brevity, we omit the subscript θ and denote the probability law $Q \in \Theta_k^\delta$, if $Q = Q_\theta$, and $\theta \in \Theta_k^\delta$.

2-B.2 Maximum likelihood (ML) estimation

Generally, suppose a probability law Q is in a certain class Ω. Then, the n-th order maximum likelihood (ML) estimator in Ω for the observed sequence z^n, is defined as

$$\hat{Q}[z^n] = \arg\max_{Q \in \Omega} Q(z^n),$$

resolving ties arbitrarily. Now, if $Q \in \Theta_k^\delta$, then there is an algorithm called expectation-maximization(EM) [4] that iteratively updates the parameter estimates to maximize the likelihood. Thus, when Q is in the class of probability laws of a HMP, the maximum likelihood estimate can be efficiently attained.[1] We denote the ML estimator in Θ_k^δ based on z^n by

$$\hat{Q}_{k,\delta}[z^n] = \arg\max_{Q \in \Theta_k^\delta} Q(z^n).$$

Obviously, when the n-tuple Z^n is random, $\hat{Q}_{k,\delta}[Z^n]$ is also a random probability law that is a function of Z^n.

[1] We neglect issues of convergence of the EM algorithm and assume that the ML estimation is performed perfectly.

2-B.3 Consistency of ML estimator

When $P_Z \in \Theta_k^\delta$, an ML estimator $\hat{Q}_{k,\delta}[Z^n]$ is said to be *strongly consistent* if

$$\lim_{n\to\infty} \hat{Q}_{k,\delta}[Z^n] = P_Z \qquad a.s.$$

The strong consistency of the ML estimator $\hat{Q}_{k,\delta}[Z^n]$ of the parameter of a finite-alphabet stationary ergodic HMP was proved in [1]. For the case of a general stationary ergodic HMP, the strong consistency was proved in [12].

We also have a sense of strong consistency for the case where P_Z is a general stationary and ergodic process. By the similar argument as in [8, Theorem 2.2.1], we have the consistency in the sense that if the observed noisy signal is not necessarily a HMP, and we still perform the ML estimation in Θ_k^δ, then we get

$$\lim_{n\to\infty} \hat{Q}_{k,\delta}[Z^n] \in \mathcal{N} \qquad a.s., \tag{1}$$

where $\mathcal{N} \triangleq \{Q \in \Theta_k^\delta : \mathbf{D}(P\|Q) = \min_{Q' \in \Theta_k^\delta} \mathbf{D}(P\|Q')\}$.[2] This second consistency result is the key result that we will use in devising and analyzing our universal filtering scheme.

3 The universal filtering problem

As mentioned in the Introduction, we will assume a stochastic setting, that is, the underlying clean signal is an output of some stationary and ergodic process whose probability law is P_X. From P_X and $\mathbf{\Pi}$, we can get the true joint probability law P and corresponding probability law of noisy observed signal, P_Z. That is,

$$\begin{aligned} P(X^n = x^n, Z^n = z^n) &= P_X(X^n = x^n) \prod_{t=1}^{n} \Pi(x_i, z_i), \quad \text{and} \\ P_Z(Z^n = z^n) &= \sum_{x^n} P(X^n = x^n, Z^n = z^n). \end{aligned}$$

A *filter* is a sequence of probability distributions $\hat{\mathbf{X}} = \{\hat{X}_t\}$, where $\hat{X}_t : \mathcal{A}^t \to \mathcal{M}$. The interpretation is that, upon observing z^t, the reconstruction for the underlying, unobserved x_t is represented by the symbol $\hat{x}$ with probability $\hat{X}_t(z^t)[\hat{x}]$. A filter is called *deterministic* if $\hat{X}_t(z^t)$ is a unit vector in $\mathbb{R}^{\mathbb{M}}$ for all t and z^t, and *randomized* if $\hat{X}_t(z^t)$ can be a simplex vector in $\mathcal{M}$ other than a unit vector for some t and z^t. The *normalized cumulative loss* of the scheme $\hat{\mathbf{X}}$ on the individual pair (x^n, z^n) is defined by

$$L_{\hat{\mathbf{X}}}(x^n, z^n) = \frac{1}{n}\sum_{t=1}^{n} \ell(x_t, \hat{X}_t(z^t)),$$

where $\ell(x_t, \hat{X}_t(z^t)) = \sum_{\hat{x}\in\hat{\mathcal{X}}} \Lambda(x_t, \hat{x})\hat{X}_t(z^t)[\hat{x}]$. Then, the goal of a filter is to minimize the *expected normalized cumulative loss* $E\Big(L_{\hat{\mathbf{X}}}(X^n, Z^n)\Big)$.

The optimal performance of the n-th order filter is defined as

$$\phi_n(P_X, \mathbf{\Pi}) = \min_{\hat{\mathbf{X}}\in\mathcal{F}} E\Big(L_{\hat{\mathbf{X}}}(X^n, Z^n)\Big),$$

where $\mathcal{F}$ denotes the class of all filters. Sub-additivity arguments similar to those in [21] imply

$$\lim_{n\to\infty} \phi_n(P_X, \mathbf{\Pi}) = \inf_{n\geq 1} \phi_n(P_X, \mathbf{\Pi}) \triangleq \mathbf{\Phi}(P_X, \mathbf{\Pi}).$$

[2] Just as in [8, Theorem 2.2.1], the notion of a.s. set convergence is used. For any subset $\mathcal{E} \in \Theta$, define $\mathcal{E}_\epsilon \triangleq \{Q \in \Theta : d(Q, \mathcal{E}) < \epsilon\}$, where d is the Euclidean distance. Then, $\lim_{n\to\infty} \hat{Q}[Z^n] \in \mathcal{E}$ a.s. if $\forall\epsilon > 0, \exists N(\epsilon, \omega)$ such that $\forall n \geq N(\epsilon, \omega), \hat{Q}[Z^n] \in \mathcal{E}_\epsilon$

By definition, $\mathbf{\Phi}(P_X,\mathbf{\Pi})$ is the (distribution-dependent) optimal asymptotic filtering performance attainable when the clean signal is generated by the law P_X and corrupted by $\mathbf{\Pi}$. This $\mathbf{\Phi}(P_X,\mathbf{\Pi})$ can be achieved by the optimal filter $\hat{\mathbf{X}}_P=\{\hat{X}_{P,t}\}$ where

$$\hat{X}_{P,t}(z^t)[\hat{x}]=Pr(B(\mathbf{P}_{X_t|z^t})=\hat{x}).$$

For brevity of notation, we denote $\hat{X}_P(z^t)=\hat{X}_{P,t}(z^t)$. Note that this is a *deterministic* filter, i.e., for a given z^t, the filter is a unit vector in $\mathbb{R}^M$ for all t. We can easily see that this filter is optimal since it minimizes $E(\ell(X_t,\hat{X}(Z^t))$ for all t, and thus, it minimizes $E\Big(L_{\hat{\mathbf{X}}}(X^n,Z^n)\Big)$ for all n.

As can be seen, $\hat{X}_P(z^t)$ needs the exact knowledge of $\mathbf{P}_{X_t|z^t}$, and thus, is dependent on the distribution of the underlying clean signal. The *universal filtering problem* is to construct (possibly a sequence of) filter(s), $\hat{\mathbf{X}}_{univ}$, that is independent of the distribution of underlying clean signal, P_X, and yet asymptotically achieving $\mathbf{\Phi}(P_X,\mathbf{\Pi})$. We describe our sequence of universal filters in the next section.

4 Universal filtering based on hidden Markov modeling

4-A Description of the filter

Before describing our sequence of universal filters, we make the following assumption on the source.

Assumption 1 *There exists a sequence of positive reals $\{\delta_k\}$, such that $\delta_k\downarrow 0$ as $k\to\infty$, and P_X satisfies*

$$P_X(X_0|X_{-k}^{-1})\geq\delta_k\quad a.s.\quad\forall k\in\mathbb{N}.\tag{2}$$

For any probability law Q, we construct a *randomized* filer as follows: For $\epsilon>0$, denote L_2 ϵ-ball in $\mathbb{R}^M$ as $B_\epsilon=\{\mathbf{V}\in\mathbb{R}^M:\|\mathbf{V}\|_2\leq\epsilon\}$. Then, we define a filter for fixed ϵ as

$$\hat{X}^\epsilon_{Q,t}(z^t)[\hat{x}]=Pr(B(\mathbf{Q}_{X_t|z^t}+\mathbf{U})=\hat{x}),\tag{3}$$

where $\mathbf{U}\in\mathbb{R}^M$ is a random vector, uniformly distributed in B_ϵ. For brevity of notation, we denote $\hat{X}^\epsilon_Q(z^t)=\hat{X}^\epsilon_{Q,t}(z^t)$. This filter is randomized since depending on Q and z^t, $\hat{X}^\epsilon_Q(z^t)$ can be a probability simplex vector in $\mathcal{M}$ that is not a unit vector. The reason we needed this randomization will be explained in proving Lemma 3.

To devise our filter, let's first consider an increasing sequence of positive integers, $\{m_i\}_{i\geq1}$, that satisfies following conditions:

$$\lim_{i\to\infty}\frac{m_{i-1}}{m_i}=0,\quad\lim_{i\to\infty}m_i=\infty.\tag{4}$$

Now, define

$$i(t)\triangleq\max\{i:m_i\leq t\}.$$

Then, given that our source distribution satisfies (2), and for fixed k, define a random probability law

$$Q_k^t\triangleq\hat{Q}_{k,\delta_k}[Z^{m_{i(t)}}]=\arg\max_{Q\in\Theta_k^{\delta_k}}Q(Z^{m_{i(t)}}).\tag{5}$$

That is, Q_k^t is the ML estimator in $\Theta_k^{\delta_k}$ based on $Z^{m_{i(t)}}$. As discussed in Section 2-B.1, we only need to estimate the state transition probabilities of the underlying Markov chain to obtain this ML estimator, and this can be efficiently done by the Expectation-Maximization (EM) algorithm. Once we get Q_k^t, we can then calculate $\mathbf{Q}^t_{kX_t|z^t}$ using the forward-recursion formula which is described in detail in [4]. Note that we get this conditional distribution directly, not by first estimating the output distribution, and then inverting the channel, as was done in [18][19][21].

Finally, we take as our sequence of universal filtering schemes, indexed by k and ϵ,

$$\hat{\mathbf{X}}^{\epsilon}_{univ,k} = \{\hat{X}^{\epsilon}_{Q^t_k,t}\}.$$

The following theorem states the main result of this paper.

Theorem 1 *Let $\mathbf{X}^\infty \in \mathcal{A}^\infty$ be a stationary, ergodic process emitted by the source P_X which satisfies Assumption 1. Let $\mathbf{Z}^\infty \in \mathcal{A}^\infty$ be the output of the DMC, $\mathbf{\Pi}$, whose input is $\mathbf{X}^\infty$. Then:*

(a) $\lim_{\epsilon\to 0}\lim_{k\to\infty}\limsup_{n\to\infty} L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n, Z^n) \le \mathbf{\Phi}(P_X, \Pi)$ *a.s.*

(b) $\lim_{\epsilon\to 0}\lim_{k\to\infty}\limsup_{n\to\infty} E\Big(L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n, Z^n)\Big) = \mathbf{\Phi}(P_X, \Pi)$

4-B Intuition behind the scheme and proof sketch

The intuition behind our scheme parallels that of the universal compression and universal prediction problems in the stochastic setting. In the n-th order problem of both cases [5][14], the excess expected codeword length per symbol and the excess expected normalized cumulative loss incurred by using the wrong probability law Q in place of the true probability law P could be upper bounded by the normalized n-th order relative entropy $\frac{1}{n}D_n(P\|Q)$. Then, to achieve the asymptotically optimum performance, the compressor and the predictor try to find and use some data-dependent Q that makes $\frac{1}{n}D_n(P\|Q) \to 0$ as $n\to\infty$, that is, makes $\mathbf{D}(P\|Q)$ zero.

We follow the same intuition in our universal filtering problem. For fixed k and ϵ, our scheme, as can be seen from (5), divides the noisy observed signal into sub-blocks of length $(m_i - m_{i-1})$. Since $\frac{m_{i-1}}{m_i}$ tends to zero as $i\to\infty$, the length of each sub-block grows faster than exponential. Now, to filter each sub-block, it plugs the ML estimator in $\Theta_k^{\delta_k}$ obtained from the entire observation of noisy signal up to the previous sub-block. From (1), we know that as the observation length n increases, this ML estimator will converge to the parameter that minimizes the relative entropy rate between the true output probability law P_Z. Then, to show that this scheme achieves the asymptotically optimum performance, we bound the excess expected normalized cumulative loss with this relative entropy rate, and show that the bound goes to zero as the HMP parameter set becomes richer, that is, k increases.

To be more specific, we briefly sketch the proof of our main theorem. Part (b) of Theorem 1 states that our scheme is asymptotically optimal. We can easily see that this follows directly from Part (a) and Reverse Fatou's Lemma. Therefore, proving Part (a) is the key in proving the theorem. Part (a) states that in the limit, the normalized cumulative loss of our scheme, for almost every realization, is less than or equal to the asymptotically optimum performance.

To prove Part (a), we first fix k and ϵ, and get the following inequality

$$\limsup_{n\to\infty}\Big(L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n, Z^n) - \phi_n(P_X, \mathbf{\Pi})\Big) \le F\Big(\limsup_{t\to\infty}\mathbf{D}(P_Z\|Q^t_k), \epsilon\Big) \qquad a.s., \tag{6}$$

where $F(x,y)$ is some function such that $F(x,y)\to 0$ as $x\downarrow 0$, and then $y\downarrow 0$.[3] There are two keys in getting this inequality. The first one is to show the concentration of $L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n, Z^n)$ to its expectation which will be shown in Lemma 3 and Corollary 1. The second is to get the explicit upper bound function $F(x,y)$ which will be based on Lemma 4. Once establishing this inequality, we show that

$$\lim_{k\to\infty}\limsup_{t\to\infty}\mathbf{D}(P_Z\|Q^t_k) = 0 \qquad a.s., \tag{7}$$

[3] Note that Q^t_k in $\mathbf{D}(P_Z\|Q^t_k)$ is a function of $Z^{m_{i(t)}}$, and thus, is random. A more formal definition of relative entropy rate between true and the random probability law like this case will be given after Lemma 4.

from Lemma 5 and then send $\epsilon \downarrow 0$ to get Part (a). Keeping this proof sketch in mind, let us move on to the detailed proof in the next section.

4-C Proof of the theorem

Before proving the theorem, we introduce several lemmas as building blocks. Lemma 1 and Lemma 2 below give some general results for the HMPs that we are considering. Our lemmas are similar to [8, Lemma 2.3.4] and [8, Theorem 2.3.3]. The latter assumed that all the parameters are lower bounded by $\delta > 0$, whereas in Θ_k^δ, some parameters can be zero. We take this into account in proving Lemma 1 and Lemma 2. Lemma 3 shows the uniform concentration property of the normalized cumulative loss on Θ_k^δ, which is an important property that we need to prove the main theorem. Lemma 4 provides a key step to get the upper bound described in (6), and Lemma 5, which needs three additional definitions, enables to show (7). After building up the lemmas, we give the proof of the main theorem, which is merely an application of the lemmas.

Lemma 1 *Suppose $Q \in \Theta_k^\delta$ and fix $\delta > 0$. Then, $\forall\omega$, $Q(Z_0|Z_{-t}^{-1})$ converges to a limit $Q(Z_0|Z_{-\infty}^{-1})$ uniformly on Θ_k^δ.*

Proof: To prove this lemma, we need three more lemmas in Appendix 1, which are variations on those found in [1]. Let's denote $f_t := Q(Z_0|Z_{-t}^{-1})$, and $f_0 = 0$. Then, the sequence $\{f_t\}$ uniformly converges on Θ_k^δ, if following k subsequences,

$$\{f_{jk+l}, j = 0, 1, 2, \cdots, \}, \qquad 0 \le l \le k-1,$$

uniformly converge on Θ_k^δ, and have the same limit.

First, the uniform convergence of each subsequence $\{f_{jk+l}\}$ can be shown by showing the series $\sum_{j=0}^t (f_{(j+1)k+l} - f_{jk+l})$ converges absolutely. From Lemma 8 in Appendix 1, setting $m = k$,

$$\begin{aligned}
&\sum_{j=0}^{t} |f_{(j+1)k+l} - f_{jk+l}| \\
=&\sum_{x_0} Q(Z_0|x_0) \sum_{j=1}^{t} |Q(x_0|Z_{-(j+1)k-l}^{-1}) - Q(x_0|Z_{-jk-l}^{-1})| \\
\le& M \sum_{j=1}^{t} (\rho_{\delta,k,k})^{j+1}.
\end{aligned}$$

Since $\rho_{\delta,k,k} < 1, M < \infty$ and $\rho_{\delta,k,k}$ does not depend on Q, ω, and l, we conclude that all k subsequences converge uniformly on Θ_k^δ.

Now, to show that the k subsequences have the same limit, construct another subsequence, $\{f_{j(k+1)+1}, j = 0, 1, 2, \cdots, \}$. Since this subsequence contains infinitely many terms from all k subsequences, if this subsequence converges uniformly on Θ_k^δ, we can conclude that the k subsequences have the same limit. The derivation of the uniform convergence of this subsequence is the same as that described above, but setting $m = k+1$ in Lemma 8. Therefore, the original sequence $\{f_t\}$ converges to its limit uniformly on Θ_k^δ. ■

The remarkable fact of this lemma is that the convergence is not only uniform on Θ_k^δ, but also in ω. That is, the convergence holds uniformly on every realization of $z_{-\infty}^0$.

Lemma 2 *For the distribution of the observed noisy process $\{Z_t\}$, P_Z, and every $Q \in \Theta_k^\delta$,*

$$\mathbf{D}(P_Z\|Q) \triangleq \lim_{n\to\infty} \frac{1}{n} D_n(P_Z\|Q) = E\Big(\log \frac{P_Z(Z_0|Z_{-\infty}^{-1})}{Q(Z_0|Z_{-\infty}^{-1})}\Big).$$

Moreover,

$$\lim_{n\to\infty}\frac{1}{n}\log\frac{P_Z(Z^n)}{Q(Z^n)} = \mathbf{D}(P_Z\|Q) \qquad \text{a.s.} \qquad \textit{uniformly on } \Theta_k^\delta.$$

Proof: This lemma consists of three parts. The first part is to show the existence of the first limit in the lemma so that the definition of $\mathbf{D}(P_Z\|Q)$ is valid. The second part is to show that the value of the limit is indeed $E\Big(\log\frac{P_Z(Z_0|Z_{-\infty}^{-1})}{Q(Z_0|Z_{-\infty}^{-1})}\Big)$. Finally, the last part is to show the uniform convergence of normalized log-likelihood ratio to the relative entropy rate. The first two parts and the pointwise convergence of the third part is a generalization of the Shannon-McMillan-Breiman theorem. The proof of these parts is identical to those in [8, Theorem 2.3.3] even for the case where some parameters in Θ_k^δ can be zero.

The uniform convergence in the third part of the lemma is crucial in that it enables to obtain the second consistency result (1) as in [8, Theorem 2.2.1]. We take into account our parameter set, and repeat the argument of [8, Lemma 2.4.1]. To show the uniform convergence, we need to show

$$\lim_{n\to\infty}\frac{1}{n}\log Q(Z^n) = E\Big(\log Q(Z_0|Z_{-\infty}^{0})\Big) \qquad a.s. \qquad \text{uniformly on } \Theta_k^\delta$$

Since the pointwise convergence can be shown and the parameter set Θ_k^δ is compact, it is enough to show that $\frac{1}{n}\log Q(Z^n)$ is an equicontinuous sequence by Ascoli's Theorem. That is, we need to show for $\forall\epsilon>0$, $\exists\delta(\epsilon)>0$ such that

$$\forall n, \left|\frac{1}{n}\log Q(Z^n) - \frac{1}{n}\log Q^{'}(Z^n)\right| \le \epsilon, \text{ if } \quad \|Q-Q^{'}\|_1 < \delta(\epsilon),$$

where $\|Q-Q^{'}\|_1 \triangleq \sum_{i,j}|a_{ij}-a^{'}_{ij}|$ is defined to be the L_1 distance between the two parameters defining Q and $Q^{'}$. This equicontinuity can be proved by observing that a process $\{S_t = (X_{t-(k-1)}^t, Z_t)\}$ is a Markov process, where $\{S_t\}$ has a state space $\mathcal{S}=\mathcal{A}^k\times\mathcal{A}$. This is true since

$$\begin{aligned}
Q(S_{t+1}|S^t) =& Q(X_{t+1}, Z_{t+1}|X^t, Z^t)\\
=& Q(X_{t+1}|X^t, Z^t)Q(Z_{t+1}|X^{t+1}, Z^t)\\
=& Q(X_{t+1}|X_{t-(k-1)}^t)\Pi(X_{t+1}, Z_{t+1})\\
=& Q(S_{t+1}|S_t).
\end{aligned}$$

Let $\{x_1^k(i): i=1,\cdots,M^k\}$ denote the set of all possible k-tuples of $\{X_t\}$, and let $s=(x_1^k(i),z), \bar{s}=(x_1^k(j),\bar{z})$. Then, the transition matrix T of $\{S_t\}$ has elements $t_{s\bar{s}} \triangleq Q(S_{t+1}=\bar{s}|S_t=s) = a_{ij}\Pi(x_k(j),\bar{z})$. Since all A that are in Θ_k^δ are irreducible and aperiodic and $\Pi(x_k(j),\bar{z})>0$, $\forall x_k(j),\bar{z}$, T is also irreducible and aperiodic. Hence, T has the unique stationary distribution τ. Although there are zeros in T, by the construction, any n-tuple s^n has positive probability. Since $\{S_t\}$ is also stationary, we have

$$Q(S^n=s^n) = \tau_{s_1}\prod_{t=k}^{n-1} t_{s_t s_{t+1}} = \tau_{s_1}\prod_{(s,\bar{s})} t_{s\bar{s}}^{n_{s\bar{s}}},$$

where

$$n_{s\bar{s}} \triangleq \sum_{t=k}^{n-1}\mathbf{1}(S_t=s, S_{t+1}=\bar{s}).$$

For another probability law $Q^{'} \in \Theta_k^{\delta}$, we have

$$
\begin{aligned}
&|\frac{1}{n}\log Q(S^n) - \frac{1}{n}\log Q^{'}(S^n)| \\
\leq &|\frac{1}{n}\log \tau_{s_1} - \frac{1}{n}\log \tau^{'}_{s_1}| + |\frac{1}{n}\sum_{(s,\bar{s})} n_{s\bar{s}} \log t_{s\bar{s}} - \frac{1}{n}\sum_{(s,\bar{s})} n_{s\bar{s}} \log t^{'}_{s\bar{s}}| \\
\leq &|\log \tau_{s_1} - \log \tau^{'}_{s_1}| + \sum_{(s,\bar{s})} |\log t_{s\bar{s}} - \log t^{'}_{s\bar{s}}| \qquad (8) \\
= &|\log \tau_{s_1} - \log \tau^{'}_{s_1}| + \sum_{(i,j)} |\log a_{ij} - \log a^{'}_{ij}| \qquad (9)
\end{aligned}
$$

where (8) is from the fact that $\frac{1}{n} \leq 1, \frac{n_{s\bar{s}}}{n} \leq 1$, and (9) is from the fact that DMC, $\mathbf{\Pi}$, is equal for Q and $Q^{'}$. The summations are over the pairs that have nonzero transition probabilities.

Since the function $f(x) = \log x$ is a uniformly continuous function for $\delta \leq x < 1$, and $a_{ij} \geq \delta$ that occur in the summation, we have for $\epsilon > 0$,

$$\sum_{(i,j)} |\log a_{ij} - \log a^{'}_{ij}| < \frac{\epsilon}{2} \quad \text{if} \quad \|Q - Q^{'}\|_1 < \delta_1(\epsilon).$$

Also, we know that all the elements of the stationary distribution of T are bounded away from zero, since the largest element of the stationary distribution of T is lower bounded by $\frac{1}{M^{k+1}}$, and any state can be reached by finite number of steps whose transition probabilities are bounded away from zero. Therefore, for some $C_1 < \infty$, we have,

$$|\log \tau_{s_1} - \log \tau^{'}_{s_1}| < C_1 |\tau_{s_1} - \tau^{'}_{s_1}|.$$

Then, from the result of the sensitivity of the stationary distribution of a Markov chain [10], for some $C_2 < \infty$, we have,

$$|\tau_{s1} - \tau^{'}_{s1}| \leq C_2 \sum_{(s,\bar{s})} |t_{s\bar{s}} - t^{'}_{s\bar{s}}| = C_2 \sum_{(i,j)} |a_{ij} - a^{'}_{ij}|.$$

Hence, for $\epsilon > 0$, we obtain,

$$|\log \tau_{s_1} - \log \tau^{'}_{s_1}| < \frac{\epsilon}{2} \quad \text{if} \quad \|Q - Q^{'}\|_1 < \delta_2(\epsilon).$$

Therefore, by letting $\delta(\epsilon) = \min(\delta_1(\epsilon), \delta_2(\epsilon))$, we have

$$\left|\frac{1}{n}\log Q(S^n) - \frac{1}{n}\log Q^{'}(S^n)\right| < \epsilon \quad \text{if} \quad \|Q - Q^{'}\|_1 < \delta(\epsilon).$$

Let us now go back to the original process Z. From

$$\left|\frac{1}{n}\log Q(S^n) - \frac{1}{n}\log Q^{'}(S^n)\right| < \epsilon,$$

we have

$$Q^{'}(X^n, Z^n) < \exp(n\epsilon) Q(X^n, Z^n),$$

thus,

$$
\begin{aligned}
Q^{'}(Z^n) &= \sum_{x^n} Q^{'}(x^n, Z^n) < \exp(n\epsilon) \sum_{x^n} Q(x^n, Z^n) \\
&= \exp(n\epsilon) Q(Z^n)
\end{aligned}
$$

where the summations are again over the sequences that have nonzero probabilities. By changing the role of Q, and $Q^{'}$, we get the result that $\frac{1}{n}\log Q(Z^n)$ is an equicontinuous sequence. Therefore, we have the uniform convergence of the lemma. ■

Lemma 3 (*Uniform Concentration*) *Suppose $Q \in \Theta_k^\delta$ for some fixed $\delta > 0$. Let $\hat{\mathbf{X}}_Q^\epsilon$ be the randomized filter defined in (3). Then,*

$$\lim_{n\to\infty} \Big(L_{\hat{\mathbf{X}}_Q^\epsilon}(X^n, Z^n) - E\Big(L_{\hat{\mathbf{X}}_Q^\epsilon}(X^n, Z^n)\Big)\Big) = 0 \quad \text{a.s.} \quad \textit{uniformly on } \Theta_k^\delta$$

Proof: This lemma shows the uniform concentration property of $L_{\hat{\mathbf{X}}_Q^\epsilon}(X^n, Z^n)$. The randomization of the filter is needed to deal with ties occur in deciding the Bayes response. A detailed proof of this lemma is given in Appendix 2.

Lemma 4 (*Continuity*) *Consider a single letter filtering setting. Suppoes Q is some other joint probability law of X and Z. Define single letter filters $\hat{X}_P(z)$ and $\hat{X}_Q^\epsilon(z)$ as*

$$\begin{aligned}\hat{X}_P(z)[\hat{x}] =& Pr(B(\mathbf{P}_{X|z}) = \hat{x})\\ \hat{X}_Q^\epsilon(z)[\hat{x}] =& Pr(B(\mathbf{Q}_{X|z} + \mathbf{U}) = \hat{x}),\end{aligned}$$

where $\mathbf{U} \in \mathbb{R}^M$ is a uniform random vector in B_ϵ as before. Then,

$$E\Big(\ell(X, \hat{X}_Q^\epsilon(Z))\Big) - E\Big(\ell(X, \hat{X}_P(Z))\Big) \le \Lambda_{\max} K_{\mathbf{\Pi}} \cdot \|\mathbf{P}_Z - \mathbf{Q}_Z\|_1 + C_{\mathbf{\Lambda}} \cdot \epsilon,$$

where the expectations on the left hand side of the inequality are under P and $K_{\mathbf{\Pi}} = \sum_{i=1}^M \|\Pi_i^{-1}\|_2$, and $C_{\mathbf{\Lambda}} = \max_{a,b\in\mathcal{A}} \|\lambda_a - \lambda_b\|_2$.

This lemma states that the excess expected loss of a randomized filter optimized for a mismatched probability law can be upper bounded by the L_1 difference between the true and the mismatched probability laws of output symbol, plus a small constant term which diminishes with the randomization probability. This is somewhat analogous to a for the prediction which was derived in [14, (20)].

Proof of Lemma 4: Define $\hat{X}_Q(z)[\hat{x}] = Pr(B(\mathbf{Q}_{X|z}) = \hat{x})$. Then,

$$\begin{aligned}
& E\Big(\ell(X, \hat{X}_Q^\epsilon(Z))\Big) - E\Big(\ell(X, \hat{X}_P(Z))\Big)\\
=& \sum_{x,z} P(x,z)\Big(\ell(x, \hat{X}_Q^\epsilon(z)) - \ell(x, \hat{X}_P(z))\Big)\\
\le& \sum_{x,z} \Big(Q(x,z) + |P(x,z) - Q(x,z)|\Big)\Big(\ell(x, \hat{X}_Q(z)) - \ell(x, \hat{X}_P(z)) + \ell(x, \hat{X}_Q^\epsilon(z)) - \ell(x, \hat{X}_Q(z))\Big)\\
\le& \sum_{x,z} |P(x,z) - Q(x,z)| \cdot \Big(\ell(x, \hat{X}_Q(z)) - \ell(x, \hat{X}_P(z))\Big) \qquad (10)\\
& + \sum_{x,z} \Big(Q(x,z) + |P(x,z) - Q(x,z)|\Big) \cdot \Big(\ell(x, \hat{X}_Q^\epsilon(z)) - \ell(x, \hat{X}_Q(z))\Big)\\
=& \sum_{x,z} |P(x,z) - Q(x,z)| \cdot \Big(\ell(x, \hat{X}_Q^\epsilon(z)) - \ell(x, \hat{X}_P(z))\Big) + \sum_{x,z} Q(x,z)\Big(\ell(x, \hat{X}_Q^\epsilon(z)) - \ell(x, \hat{X}_Q(z))\Big) \qquad (11)\\
\le& \Lambda_{max} \sum_{x,z} |P(x,z) - Q(x,z)| + \sum_{x,z} Q(x,z)\Big(\ell(x, \hat{X}_Q^\epsilon(z)) - \ell(x, \hat{X}_Q(z))\Big), \qquad (12)
\end{aligned}$$

where (10) is from the fact that $\sum_{x,z} Q(x,z)(\ell(x,\hat{X}_Q(z)) - \ell(x,\hat{X}_P(z))) \leq 0$ and (11) is from rearranging terms in the summation. Now, let's bound the first term in (12).

$$
\begin{aligned}
&\Lambda_{\max} \sum_{x,z} |P(x,z) - Q(x,z)| \\
=&\Lambda_{\max} \sum_{x} |P(x) - Q(x)| \Big(\sum_{z} \Pi(x,z) \Big) \\
=&\Lambda_{\max} \sum_{x} |P(x) - Q(x)| && (13)\\
=&\Lambda_{\max} \sum_{i} |(\mathbf{P}_Z - \mathbf{Q}_Z)^T \Pi_i^{-1}| \\
\leq&\Lambda_{\max} \sum_{i} \|\Pi_i^{-1}\|_2 \cdot \|\mathbf{P}_Z - \mathbf{Q}_Z\|_2 && (14)\\
\leq&\Lambda_{\max} K_{\mathbf{\Pi}} \cdot \|\mathbf{P}_Z - \mathbf{Q}_Z\|_1, && (15)
\end{aligned}
$$

where (13) is from the fact that $\sum_z \Pi(x,z) = 1$, (14) is from Cauchy-Schwartz inequality, and (15) is from the fact that L_2-norm is less than or equal to L_1-norm.

The second term in (12) becomes

$$
\begin{aligned}
&\sum_{x,z} Q(x,z) \Big(\ell(x, \hat{X}_Q^{\epsilon}(z)) - \ell(x, \hat{X}_Q(z)) \Big) \\
=&\sum_{z} Q(z) \sum_{x} Q(x|z) \sum_{\hat{x}} \Lambda(x,\hat{x}) \cdot \Big(\hat{X}_Q^{\epsilon}(z)[\hat{x}] - \hat{X}_Q(z)[\hat{x}] \Big) \\
=&\sum_{z} Q(z) \sum_{\hat{x}} \Big(\hat{X}_Q^{\epsilon}(z)[\hat{x}] - \hat{X}_Q(z)[\hat{x}] \Big) \sum_{x} \Lambda(x,\hat{x}) Q(x|z) \\
=&\sum_{z} Q(z) \sum_{\hat{x}} \Big(\hat{X}_Q^{\epsilon}(z)[\hat{x}] - \hat{X}_Q(z)[\hat{x}] \Big) \cdot \lambda_{\hat{x}}^T \mathbf{Q}_{X|z}. && (16)
\end{aligned}
$$

It is easy to see that the inner summation in (16) is always nonnegative since by definition, $\hat{X}_Q(z)$ assigns probability 1 to $B(\mathbf{Q}_{X|z})$. Now, for a given Q, define

$$
\mathbf{U}_{\max} = \arg\max_{\mathbf{U} \in B_{\epsilon}} \Big(\lambda_{B(\mathbf{Q}_{X|z} + \mathbf{U})} - \lambda_{B(\mathbf{Q}_{X|z})} \Big)^T \mathbf{Q}_{X|z}, \tag{17}
$$

resolving ties arbitrarily. Then, we have,

$$
\begin{aligned}
&\sum_{\hat{x}} \Big(\hat{X}_Q^{\epsilon}(z)[\hat{x}] - \hat{X}_Q(z)[\hat{x}] \Big) \cdot \lambda_{\hat{x}}^T \mathbf{Q}_{X|z} \\
=&\Big(\sum_{\hat{x}} \Big(\hat{X}_Q^{\epsilon}(z)[\hat{x}] \cdot \lambda_{\hat{x}} \Big) - \lambda_{B(\mathbf{Q}(X|z))} \Big)^T \mathbf{Q}_{X|z} \\
\leq&\Big(\lambda_{B(\mathbf{Q}(X|z) + \mathbf{U}_{\max})} - \lambda_{B(\mathbf{Q}(X|z))} \Big)^T \mathbf{Q}_{X|z} && (18)\\
\leq&\Big(\lambda_{B(\mathbf{Q}(X|z))} - \lambda_{B(\mathbf{Q}_{X|z} + \mathbf{U}_{\max})} \Big)^T \mathbf{U}_{\max} && (19)\\
\leq& \max_{a,b \in \mathcal{A}} \|\lambda_a - \lambda_b\|_2 \cdot \|\mathbf{U}_{\max}\|_2 && (20)\\
\leq& C_{\mathbf{\Lambda}} \cdot \epsilon,
\end{aligned}
$$

where (18) follows from (17), (19) follows from the fact

$$
\lambda_{B(\mathbf{Q}_{X|z} + \mathbf{U}_{\max})}^T (\mathbf{Q}_{X|z} + \mathbf{U}_{\max}) \leq \lambda_{B(\mathbf{Q}_{X|z})}^T (\mathbf{Q}_{X|z} + \mathbf{U}_{\max}),
$$

and (20) follows from the Cauchy-Schwartz inequality. Note that depending on Q and z, (18) and (19) can be both zero and hold with equality. Together with (15), the lemma is proved. ■

Before moving on to Lemma 5, we need following three definitions. In Lemma 2, we have seen that for $Q \in \Theta_k^\delta$, $\mathbf{D}(P_Z\|Q)$ is well-defined. Now, let's consider the case where $Q \in \Theta_k^\delta$ is some function of the noisy observation Z^n (denoted as $Q[Z^n]$). As mentioned in the footnote of Section 4-B, the notion of the relative entropy rate between P_Z and that random $Q[Z^n]$ is defined in Definition 2 using Definition 1. Definition 3 is also needed for the inequality in Lemma 5.

Definition 1 *Suppose $Q[Z^n] \in \Theta_k^\delta$. If f is some function of $(X^\infty, Z^\infty, Q[Z^n])$ such that the expectation*

$$E\Big(f(X^\infty, Z^\infty, Q[Z^n])\Big) = \int f(x^\infty, z^\infty, Q[z^n])dP(x^\infty, z^\infty)$$

exists. Then, define the notation $\hat{E}(\cdot)$ as following:

$$\hat{E}\Big(f(X^\infty, Z^\infty, Q[Z^n])\Big) \triangleq \int f(x^\infty, z^\infty, Q[Z^n])dP(x^\infty, z^\infty)$$

That is, in $\hat{E}\Big(f(X^\infty, Z^\infty, Q[Z^n])\Big)$, the Lebesgue integration with respect to the randomness of $Q[Z^n]$ is excluded.

Definition 2 *Suppose $Q[Z^n] \in \Theta_k^\delta$. Then, the relative entropy rate between P_Z and $Q[Z^n]$ is defined as,*[4]

$$\mathbf{D}(P_Z\|Q[Z^n]) \triangleq \hat{E}\Big(\log\frac{P_Z(Z_0|Z_{-\infty}^{-1})}{Q[Z^n](Z_0|Z_{-\infty}^{-1})}\Big).$$

Definition 3 *Define the k-th order Markov approximation of P_X for $n \geq k$ as*

$$P_X^{(k)}(X^n) \triangleq P_X(X^k)\prod_{i=k+1}^{n} P_X(X_i|X_{i-k}^{i-1}).$$

Furthermore, denote P_Z and $P_Z^{(k)}$ as the probability law of the output of DMC, $\mathbf{\Pi}$, when the probability law of input is P_X and $P_X^{(k)}$, respectively.[5]

Now, we give following lemma that upper bounds the relative entropy rate between P_Z and the ML estimator.

Lemma 5 *For the given sequence $\{\delta_k\}$ defined in Section 4-A and for fixed k, we have*

$$\lim_{n\to\infty}\mathbf{D}(P_Z\|\hat{Q}_{k,\delta_k}[Z^n]) \leq \mathbf{D}(P_X\|P_X^{(k)}) \quad \text{a.s.}$$

Proof: Recall that $\hat{Q}_{k,\delta_k}[Z^n]$ is an ML estimator in $\Theta_k^{\delta_k}$ based on the observation Z^n. From (1), we know that

$$\lim_{n\to\infty}\mathbf{D}(P_Z\|\hat{Q}_{k,\delta_k}[Z^n]) = \min_{Q\in\Theta_k^{\delta_k}}\mathbf{D}(P_Z\|Q) \quad \textit{a.s.}$$

Also, (2) and Definition 3 assures that $P_Z^{(k)} \in \Theta_k^{\delta_k}$. Therefore, we have

$$\lim_{n\to\infty}\mathbf{D}(P_Z\|\hat{Q}_{k,\delta_k}[Z^n]) \leq \mathbf{D}(P_Z\|P_Z^{(k)}) \quad \textit{a.s.}$$

[4]Note that $\mathbf{D}(P_Z\|Q[Z^n])$ is a function of Z^n, and still is a random variable.

[5]Here, $P_Z^{(k)}$ is not the k-th order Markov approximation of P_Z, but is the distribution of the channel output whose input is $P_X^{(k)}$, the k-th order Markov approximation of the original input distribution P_X.

This is the link where we needed Assumption 1. Now, let's denote $P^{(k)}$ as the joint probability law of (X^n, Z^n) when the probability law of input process is $P_X^{(k)}$. Then, by the chain rule of relative entropy [5, (2.67)], we have

$$\begin{aligned}
&E\Big(\log\frac{P(X^n,Z^n)}{P^{(k)}(X^n,Z^n)}\Big)\\
=&D_n(P_X\|P_X^{(k)})+E\Big(\log\frac{P(Z^n|X^n)}{P^{(k)}(Z^n|X^n)}\Big)\\
=&D_n(P_Z\|P_Z^{(k)})+E\Big(\log\frac{P(X^n|Z^n)}{P^{(k)}(X^n|Z^n)}\Big)
\end{aligned}$$

Since the DMC is fixed, we have $E\Big(\log\frac{P(Z^n|X^n)}{P^{(k)}(Z^n|X^n)}\Big)=0$. Moreover, by the nonnegativity of relative entropy, $E\Big(\log\frac{P(X^n|Z^n)}{P^{(k)}(X^n|Z^n)}\Big)\geq 0$. Therefore, we get $D_n(P_Z\|P_Z^{(k)})\leq D_n(P_X\|P_X^{(k)})$. Since $\mathbf{D}(P_X\|P_X^{(k)})=\lim_{n\to\infty}\frac{1}{n}D_n(P_X\|P_X^{(k)})$ always exists by ergodicity, we have

$$\mathbf{D}(P_Z\|P_Z^{(k)})\leq\mathbf{D}(P_X\|P_X^{(k)}),$$

and the lemma is proved. ■

Proof of Theorem 1 We are now finally in a position to prove our main theorem. As mentioned in Section 4-B, we first fix k and ϵ, and try to get the inequality in the form of (6) to prove Part (a). To refresh, (6) is given again here.

$$\limsup_{n\to\infty}\Big(L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)-\phi_n(P_X,\mathbf{\Pi})\Big)\leq F\Big(\limsup_{t\to\infty}\mathbf{D}(P_Z\|Q_k^t),\epsilon\Big)\qquad a.s.$$

From the definition of $L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)$,

$$L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)=\frac{1}{n}\sum_{t=1}^n\ell(X_t,\hat{X}^\epsilon_{Q_k^t}(Z^t)),$$

where from (5), we know that Q_k^t is a function of $Z^{m_i(t)}$. Since $\ell(X_t,\hat{X}^\epsilon_{Q_k^t}(Z^t))$ is a function of $(X_t,Z^t,Q[Z^{m_i(t)}])$, we can define a quantity $\hat{E}(\ell(X_t,\hat{X}^\epsilon_{Q_k^t}(Z^t)))$ from Definition 1. From this, we also define

$$\hat{E}\Big(L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)\Big)=\frac{1}{n}\sum_{t=1}^n\hat{E}\Big(\ell(X_t,\hat{X}^\epsilon_{Q_k^t}(Z^t))\Big).$$

Now, we have following Corollary 1 from Lemma 3, whose proof is given in Appendix 3. This corollary is a key step in proving the main theorem, since it provides a crucial link that enables to get the inequality in (6).

Corollary 1 *For fixed k and ϵ, we have*

$$\lim_{n\to\infty}\Big(L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)-\hat{E}\Big(L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)\Big)\Big)=0\quad a.s.$$

From Corollary 1, we have following equality

$$\begin{aligned}
&\limsup_{n\to\infty}\Big(L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)-\phi_n(P_X,\Pi)\Big)\\
=&\limsup_{n\to\infty}\Big(\hat{E}\Big(L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)\Big)-\phi_n(P_X,\Pi)\Big)\quad a.s.
\end{aligned}$$

Therefore , to get the inequality of the form of (6), we can equivalently show

$$\limsup_{n\to\infty}\Big(\hat{E}\Big(L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)\Big)-\phi_n(P_X,\Pi)\Big)\leq F\Big(\limsup_{t\to\infty}\mathbf{D}(P_Z\|Q_k^t),\epsilon\Big).$$

Now, let's consider following chain of inequalities:

$$
\begin{aligned}
&\hat{E}\Big(L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n,Z^n)\Big)-\phi_n(P_X,\Pi)\\
=&\frac{1}{n}\sum_{t=1}^{n}\Big(\hat{E}\Big(\ell(X_t,\hat{X}^{\epsilon}_{Q^t_k}(Z^t))\Big)-\hat{E}\Big(\ell(X_t,\hat{X}_P(Z^t))\Big)\Big)\quad a.s.\\
=&\frac{1}{n}\sum_{t=1}^{n}\hat{E}\Big(\hat{E}\Big(\ell(X_t,\hat{X}^{\epsilon}_{Q^t_k}(Z_t,Z^{t-1}))|Z^{t-1}\Big)-\hat{E}\Big(\ell(X_t,\hat{X}_P(Z_t,Z^{t-1}))|Z^{t-1}\Big)\Big)\quad a.s.\\
\le&\frac{K_{\mathbf{\Pi}}\Lambda_{\max}}{n}\sum_{t=1}^{n}\hat{E}\|\mathbf{P}_{Z_t|Z^{t-1}}-\mathbf{Q^t_k}_{Z_t|Z^{t-1}}\|_1+C_{\mathbf{\Lambda}}\cdot\epsilon\quad a.s. &(21)\\
\le&\frac{\sqrt{2\ln 2}K_{\mathbf{\Pi}}\Lambda_{\max}}{n}\sum_{t=1}^{n}\hat{E}\sqrt{\hat{E}\Big(\log\frac{P(Z_t|Z^{t-1})}{Q^t_k(Z_t|Z^{t-1})}\Big|Z^{t-1}\Big)}+C_{\mathbf{\Lambda}}\cdot\epsilon\quad a.s. &(22)\\
\le&\sqrt{2\ln 2}K_{\mathbf{\Pi}}\Lambda_{\max}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\hat{E}\Big(\log\frac{P(Z_t|Z^{t-1})}{Q^t_k(Z_t|Z^{t-1})}\Big)}+C_{\mathbf{\Lambda}}\cdot\epsilon.\quad a.s. &(23)
\end{aligned}
$$

(21) is obtained from Lemma 4, since $\mathbf{\Pi}$ does not vary with t, and given Z^{t-1}, estimating X_t based on Z^t is equivalent to the single letter setting as in Lemma 4 with the corresponding conditional distribution. Also, (22) is from Pinsker's inequality, and (23) is from Jensen's inequality. By taking $\limsup$ on both sides, we have

$$
\begin{aligned}
&\limsup_{n\to\infty}\Big(\hat{E}\Big(L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n,Z^n)\Big)-\phi_n(P_X,\Pi)\Big)\\
\le&\sqrt{2\ln 2}K_{\mathbf{\Pi}}\Lambda_{\max}\sqrt{\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\hat{E}\Big(\log\frac{P(Z_t|Z^{t-1})}{Q^t_k(Z_t|Z^{t-1})}\Big)}+C_{\mathbf{\Lambda}}\cdot\epsilon\quad a.s.
\end{aligned}
$$

since the square root function is a continuous function. For the expression inside the square root of the right-hand side of the inequality,

$$
\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\hat{E}\Big(\log\frac{P(Z_t|Z^{t-1})}{Q^t_k(Z_t|Z^{t-1})}\Big)\\
=&\limsup_{t\to\infty}\hat{E}\Big(\log\frac{P(Z_t|Z^{t-1})}{Q^t_k(Z_t|Z^{t-1})}\Big)\quad a.s. &(24)\\
=&\limsup_{t\to\infty}\hat{E}\Big(\log\frac{P(Z_0|Z^{-1}_{-\infty})}{Q^t_k(Z_0|Z^{-1}_{-\infty})}\Big)\quad a.s. &(25)\\
=&\limsup_{t\to\infty}\mathbf{D}(P_Z\|Q^t_k)\quad a.s. &(26)
\end{aligned}
$$

where (24) is from Cesáro's mean convergence theorem; (25) is from the fact that $P(Z_0|Z^{-1}_{-t})\to P(Z_0|Z^{-1}_{-\infty})$ almost surely by martingale convergence theorem, and $Q^t_k(Z_t|Z^{t-1})\to Q^t_k(Z_0|Z^{-1}_{-\infty})$ almost surely by Lemma 1, and (26) is from Definition 2. Therefore,

$$
\begin{aligned}
&\limsup_{n\to\infty}\Big(\hat{E}\Big(L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n,Z^n)\Big)-\phi_n(P_X,\Pi)\Big)\\
&\limsup_{n\to\infty}\Big(L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n,Z^n)-\phi_n(P_X,\Pi)\Big)\\
\le&2\sqrt{2\ln 2}K_{\Pi}\Lambda_{max}\sqrt{\limsup_{t\to\infty}\mathbf{D}(P_Z\|Q^t_k)}+C_{\mathbf{\Lambda}}\cdot\epsilon\quad a.s. &(27)
\end{aligned}
$$

which finally is in the form of (6). Now, we need to check if the right-hand side of (27) goes to zero if we let $k \to \infty$ and $\epsilon \downarrow 0$. To see this, consider following further upper bounds.

$$
\begin{aligned}
&\limsup_{t\to\infty} \mathbf{D}(P_Z \| Q_k^t) \\
=&\limsup_{t\to\infty} \mathbf{D}(P_Z \| \hat{Q}_{k,\delta_k}[Z^t]) && (28)\\
\le&\mathbf{D}(P_X \| P_{X_k}), && (29)
\end{aligned}
$$

where (28) is from the fact that $m_{i(t)} \to \infty$ as $t \to \infty$, and (29) is from Lemma 5. The inequality (29) holds for every k, and by Shannon-McMillan-Breiman Theorem, we know $\mathbf{D}(P_X \| P_{X_k}) \to 0$ as $k \to \infty$. Therefore,

$$\lim_{k\to\infty} \limsup_{t\to\infty} \mathbf{D}(P_Z \| Q_k^t) = 0,$$

and thus,

$$\lim_{k\to\infty} \limsup_{n\to\infty} \Big(L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n, Z^n) - \phi_n(P_X, \Pi) \Big) \le C_{\mathbf{\Lambda}} \cdot \epsilon \quad a.s.$$

Finally, sending ϵ to zero, Part (a) of the theorem is proved. Part (b) follows directly from (a), and Reverse Fatou's Lemma. That is,

$$
\begin{aligned}
&\lim_{k\to\infty} \limsup_{n\to\infty} \Big(E\Big(L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n, Z^n) \Big) - \phi_n(P_X, \Pi) \Big) \\
=&\lim_{k\to\infty} \limsup_{n\to\infty} E\Big(L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n, Z^n) - \phi_n(P_X, \Pi) \Big) \\
\le&\lim_{k\to\infty} E\left(\limsup_{n\to\infty} \Big(L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n, Z^n) - \phi_n(P_X, \Pi) \Big) \right) \\
\le&C_{\mathbf{\Lambda}} \cdot \epsilon
\end{aligned}
$$

Note that the expectation here is with respect to the randomness of probability law within the paranthesis, too. By sending ϵ to zero, Part (b) is proved. ■

5 Extension: Universal filtering for channel with memory

Now, let's extend our result to the case where channel has memory. With the identical assumption on $\{X_t\}$, now suppose $\{Z_t\}$ is expressed as

$$Z_t = X_t \oplus N_t \tag{30}$$

where $\oplus$ denotes modulo-M addition, and $\{N_t\}$ is an $\mathcal{A}$-valued noise process which is not necessarily memoryless. We assume we have a complete knowledge of the probability law of $\{N_t\}$. Specifically, let's consider the case where $\{N_t\}$ is FS-HMP, that is, it is an output of an invertible memoryless channel $\mathbf{\Gamma} = \{\Gamma(i,j)\}_{i,j\in\mathcal{A}}$ whose input is irreducible, aperiodic ℓ-th order Markov chain, $\{S_t\}$, which is independent of $\{X_t\}$. Let $\Gamma_{\min} = \min_{i,j\in\mathcal{A}}\{\Gamma(i,j)\}$, and suppose $\Gamma_{\min} > 0$. For simplicity, assume that the alphabet size of $\{S_t\}$ is also $\mathcal{A}$.

In this model, the channel between X_t and Z_t at time t is an M-ary symmetric channel, which is specified by the S_t-th row of $\mathbf{\Gamma}$. Let's define an $M \times M$ matrix $\mathbf{\Pi}_t$ whose (x_t, z_t)-th element is

$$
\begin{aligned}
\Pi_t(x_t, z_t) =&P_{N_t}(z_t \ominus x_t) \\
=&Pr(Z_t = z_t | X_t = x_t) \\
=&\sum_{s_t} Pr(Z_t = z_t | X_t = x_t, S_t = s_t) Pr(S_t = s_t),
\end{aligned}
$$

where $\ominus$ denotes modulo-M subtraction. Now, let's make following assumptions on the noise process.

- $\{N_t\}$ is stationary, i.e., $\mathbf{\Pi}_t$ is identical for $\forall t$
- $\mathbf{\Pi}_t$ is invertible
- $\exists \alpha$ such that $\quad Pr(S_t|S_{t-\ell}^{t-1}) \geq \alpha > 0, \quad \text{for} \quad \forall S_{t-\ell}^t(\omega)$

As stated in [22, 2-A], the first and the second assumptions are rather benign. Especially, for the second assumption, it can be shown that under benign conditions on the parametrization, almost all parameter values except for those in a set of Lebesgue measure zero, give rise to a process satisfying this assumption. Also, since this only corresponds to the case when $k = 0$ in [22, Assumption 1], it is a much weaker assumption. The third assumption is a similar positivity assumption as Assumption 1, which enables our universal filtering scheme.

Under these assumptions on the noise process, we can extend our scheme to do the universal filtering for this channel. First, we can convert this channel to the equivalent memoryless channel, $\mathbf{\Xi} = \{\xi((i,j),h)\}_{i,j,h\in\mathcal{A}}$, where the input process is $\{(X_t, S_t)\}$ and the output is $\{Z_t\}$. Here, $\mathbf{\Xi}$ is $M^2 \times M$ matrix, and the channel transition probability is

$$\xi((i,j),h) = \Gamma(j, h \ominus i) \quad \forall i,j,k.$$

To do the filtering, we apply our scheme to this equivalent memoryless channel. For fixed $k \geq \ell$, as in Section 2-B.1, define a parameter set of HMPs, Θ_k, whose Markov chain has $M^{k+\ell}$ states, and the memoryless channel is $\mathbf{\Xi}$. The k-th order conditional probability of our new input process is

$$\begin{aligned}
&Pr(X_t, S_t|X_{t-k}^{t-1}, S_{t-k}^{t-1}) \\
=&Pr(X_t|X_{t-k}^{t-1}) \cdot Pr(S_t|S_{t-\ell}^{t-1}) \\
\geq&\delta_k \cdot \alpha. \qquad (31)
\end{aligned}$$

where (31) is from Assumption 1 and the third condition on the noise process. Let $\gamma_k = \delta_k \cdot \alpha$. Then, we can model $\{Z_t\}$ in $\Theta_k^{\gamma_k}$, or equivalently, model (X_t, S_t) as k-th order Markov chain, and obtain Q_k^t, the ML estimator in $\Theta_k^{\gamma_k}$ based on $Z^{m_{i(t)}}$. By forward recursion, we can get $Q_k^t(X_t, S_t|Z^t)$, and by summing over S_t's we can calculate $\mathbf{Q}_{kX_t|Z^t}^t$. Then, finally we define our sequence of universal filtering schemes as,

$$\hat{\mathbf{X}}_{univ,k}^{\epsilon} = \{\hat{X}_{Q_k^t,t}^{\epsilon}\},$$

exactly the same as we proposed in Section 4-A.

The analysis of this scheme is identical to the one given in the proof of the main theorem. (21), which is the only place where the invertibility of the $\mathbf{\Pi}$ is used, can also be obtained in this case due to the second assumption of the noise process. Thus, we again get

$$\begin{aligned}
&\limsup_{n\to\infty} \Big(L_{\hat{\mathbf{X}}_{univ,k}^{\epsilon}}(X^n, Z^n) - \phi_n(P_X, \Pi) \Big) \\
\leq& 2\sqrt{2\ln 2} K_{\Pi} \Lambda_{max} \sqrt{\limsup_{t\to\infty} \mathbf{D}(P_Z \| Q_k^t)} + C_{\mathbf{\Lambda}} \cdot \epsilon \quad a.s.
\end{aligned}$$

Since

$$\limsup_{t\to\infty} \mathbf{D}(P_Z \| Q_k^t) = \limsup_{t\to\infty} \mathbf{D}(P_Z \| \hat{Q}_{k,\gamma_k}[Z^t]) \leq \mathbf{D}(P_X \| P_{X_k})$$

by the same argument as Lemma 5, we have the same result as Theorem 1. Thus, we can successfully extend our scheme to the case where the channel noise is FS-HMP with some mild assumptions.

6 Conclusion and future work

In this paper, we proved that, for the known, invertible DMC, a family of filters based on HMPs is universally asymptotically optimal for any general stationary and ergodic $\{X_t\}$ satisfying some mild positivity condition. That is, we showed that our sequence of schemes indexed by k and ϵ achieves the best asymptotically optimal performance regardless of clean source distribution. We could also extend this scheme to the case where channel has memory, especially where the channel noise process is FS-HMP. The future direction of the work would be to ascertain the relationship between k and n, such that we can devise a single scheme that grows k with some rate related to n. Attempting to loosen the positivity assumption that we made in our main theorem and extending our discrete universal filtering schemes to discrete universal denoising schemes are additional future directions of our research.

Acknowledgement

The authors are grateful to Erik Ordentlich and Sergio Verdú for helpful discussions.

Appendix 1

Here, we revise three lemmas from [1] regarding probability law of HMP. These are needed to prove Lemma 1. For the following three lemmas, fix k and δ, and suppose $Q \in \Theta_k^\delta$. Also, fix some $m \in \mathbb{N}$, such that $m \geq k$. Proofs are similar to [1, Appendix]. Note that $\{X_t\}$ is still our clean signal and $\{Z_t\}$ is the noisy observed signal (not necessarily a HMP).

Lemma 6 We have

$$Q(X_{t+m} = j | X_t = i, Z_{-\infty}^{\infty}) \geq \mu_{\delta,k,m},$$

where $\mu_{\delta,m,k} = (1 + \frac{M-1}{(\delta \cdot \Pi_{\min})^{m+k}})^{-1}$ is independent of $Q, Z_{-\infty}^{\infty}, i, j$.

Proof:

$$\begin{aligned}
&\frac{Q(X_{t+m} = j | X_t = i, Z_{-\infty}^{\infty})}{Q(X_{t+m} = j' | X_t = i, Z_{-\infty}^{\infty})} \\
=&\frac{Q(X_{t+m} = j, Z_{-\infty}^{\infty} | X_t = i)}{Q(X_{t+m} = j', Z_{-\infty}^{\infty} | X_t = i)} \\
=&\frac{Q(X_{t+m} = j, Z_{t+m+k+1}^{\infty} | X_t = i)}{Q(X_{t+m} = j', Z_{t+m+k+1}^{\infty} | X_t = i)} \cdot \frac{Q(Z_{t+1}^{t+m+k} | X_t = i, X_{t+m} = j)}{Q(Z_{t+1}^{t+m+k} | X_t = i, X_{t+m} = j')}
\end{aligned} \tag{32}$$

Now, let's bound the terms in (32). First,

$$\begin{aligned}
&\frac{Q(X_{t+m} = j, Z_{t+m+k+1}^{\infty} | X_t = i)}{Q(X_{t+m} = j', Z_{t+m+k+1}^{\infty} | X_t = i)} \\
=&\frac{\sum_{j_0} Q(X_{t+m+k} = j_0, X_{t+m} = j, Z_{t+m+k+1}^{\infty} | X_t = i)}{\sum_{j_0} Q(X_{t+m+k} = j_0, X_{t+m} = j', Z_{t+m+k+1}^{\infty} | X_t = i)} \\
=&\frac{\sum_{j_0} a_{ij}^m a_{jj_0}^k Q(Z_{t+m+k+1}^{\infty} | X_{t+m+k} = j_0)}{\sum_{j_0} a_{ij'}^m a_{j'j_0}^k Q(Z_{t+m+k+1}^{\infty} | X_{t+m+k} = j_0)}.
\end{aligned}$$

Note that $a_{ij}^m \geq \delta^m$ and $a_{jj_0}^k \geq \delta^k$, $\forall i, j, j_0$ from the assumption of Θ_k^δ. Let $Q(Z_{t+m+k+1}^\infty | X_{t+m+k} = j_0) = \alpha_{j_0}$. Then, the last expression is

$$\frac{a_{ij}^m}{a_{ij'}^m} \frac{\sum_{j_0} a_{jj_0}^k \alpha_{j_0}}{\sum_{j_0} a_{j'j_0}^k \alpha_{j_0}}. \tag{33}$$

Since

$$\frac{\sum_{j_0} a_{jj_0}^k \alpha_{j_0}}{\sum_{j_0} a_{j'j_0}^k \alpha_{j_0}} = \frac{\sum_{j_0} \alpha_{j_0} a_{j'j_0}^k \frac{a_{jj_0}^k}{a_{j'j_0}^k}}{\sum_{j_0} \alpha_{j_0} a_{j'j_0}^k} \leq \max_{j_0} \Big(\frac{a_{jj_0}^k}{a_{j'j_0}^k} \Big),$$

we have

$$(33) \leq \frac{a_{ij}^m}{a_{ij'}^m} \max_{j_0} \Big(\frac{a_{jj_0}^k}{a_{j'j_0}^k} \Big) \leq \max_{i,j,j',j_0} \Big(\frac{a_{ij}^m a_{jj_0}^k}{a_{ij'}^m a_{j'j_0}^k} \Big) \leq \frac{1}{\delta^{m+k}}. \tag{34}$$

Now let's look at the second term in (32). That is,

$$\begin{aligned}
&\frac{Q(Z_{t+1}^{t+m+k} | X_t = i, X_{t+m} = j)}{Q(Z_{t+1}^{t+m+k} | X_t = i, X_{t+m} = j')} \\
=&\frac{\sum_{x_\mathcal{T}} Q(Z_{t+1}^{t+m+k} | X_t = i, X_{t+m} = j, X_\mathcal{T} = x_\mathcal{T}) \cdot Q(X_\mathcal{T} = x_\mathcal{T} | X_t = i, X_{t+m} = j)}{\sum_{x_\mathcal{T}} Q(Z_{t+1}^{t+m+k} | X_t = i, X_{t+m} = j', X_\mathcal{T} = x_\mathcal{T}) \cdot Q(X_\mathcal{T} = x_\mathcal{T} | X_t = i, X_{t+m} = j')} \\
\leq&\frac{1}{(\Pi_{\min})^{m+k}}
\end{aligned} \tag{35}$$

where $\mathcal{T} = \{t+1, \cdots, t+m+k\} \backslash \{t, t+m\}$. Thus, from (34) and (35),

$$(32) \leq \frac{1}{(\delta \cdot \Pi_{\min})^{m+k}}.$$

Let now $\rho_j \triangleq Q(X_{t+m} = j | X_t = i, Z_{-\infty}^\infty)$, then $1 = \rho_j + \sum_{j' \neq j} \rho_{j'} \leq \rho_j + (M-1) \frac{\rho_j}{(\delta \cdot \Pi_{\min})^{m+k}}$, and thus, $\rho_j \geq (1 + \frac{M-1}{(\delta \cdot \Pi_{\min})^{m+k}})^{-1}$, which proves the lemma.

Lemma 7 Consider following two arbitrarily given sets.

$$\begin{aligned}
C_t &\in \mathcal{X}_t^\infty \triangleq \Big\{ x_\mathcal{T} : \mathcal{T} \subseteq \mathbb{Z}_{\geq t} \cup \{\infty\} \Big\} \quad \text{and} \\
D &\in \mathcal{Z}_{-\infty}^\infty \triangleq \Big\{ z_\mathcal{T} : \mathcal{T} \subseteq \mathbb{Z} \cup \{\infty, -\infty\} \Big\}.
\end{aligned}$$

For $d \in \mathbb{N}$, define

$$\begin{aligned}
M_d^+ &\triangleq \max_i Q(C_t | X_{t-dm} = i, D), \\
M_d^- &\triangleq \min_i Q(C_t | X_{t-dm} = i, D).
\end{aligned}$$

Then,

$$M_d^+ - M_d^- \leq (\rho_{\delta,k,m})^{d-1}$$

where $\rho_{\delta,k,m} = 1 - 2\mu_{\delta,k,m}$.

Proof: From the argument of Lemma 6, it is easy to see that

$$Q(X_{t+m} = j | X_t = i, D) \geq \mu_{\delta,k,m},$$

independent of D, too. Now, define

$$\begin{aligned}
\gamma_i(d) &\triangleq Q(C_t|X_{t-dm}=i,D)\\
\beta_{ij}(d) &\triangleq Q(X_{t-dm}=j|X_{t-(d+1)m}=i,D)\\
i^+(d) &\triangleq \arg\max_i Q(C_t|X_{t-(d+1)m}=i,D)\\
i^-(d) &\triangleq \arg\min_i Q(C_t|X_{t-dm}=i,D).
\end{aligned}$$

Since δ,k and m are fixed, let's simply denote $\mu=\mu_{\delta,k,m}$. Also, let's omit d and the parenthesis for above four quantities to simplify notation. Then,

$$\begin{aligned}
M_{d+1}^+ =& Q(C_t|X_{t-(d+1)m}=i^+,D)=\sum_j \gamma_j\beta_{i^+j}\\
=& \mu M_d^- + (\beta_{i^+i^-}-\mu)M_d^- + \sum_{j\neq i^-}\gamma_j\beta_{i^+j} &(36)\\
\leq& \mu M_d^- + (\beta_{i^+i^-}-\mu)M_d^+ + \sum_{j\neq i^-}\beta_{i^+j}M_d^+\\
=& \mu M_d^- + (1-\mu)M_d^+ &(37)
\end{aligned}$$

where (36) is possible from Lemma 6, since $\beta_{ij}\geq\mu$ for $\forall i,j$.
By the similar argument, we get

$$M_{d+1}^- \geq \mu M_d^+ + (1-\mu)M_d^- \tag{38}$$

By subtracting (38) from (37), we get

$$M_{d+1}^+ - M_{d+1}^- \leq (1-2\mu)(M_d^+ - M_d^-) \leq \cdots \leq (1-2\mu)^d$$

and, thus proves the lemma. Note that since $\mu=\mu_{\delta,k,m}<\frac{1}{2}$, and thus, $0<\rho_{\delta,k,m}<1$. Also, the result does not depend on Q.

Lemma 8

$$|Q(C_t|Z_{t-dm-l}^p) - Q(C_t|Z_{t-(d+1)m-l}^p)| \leq (\rho_{\delta,k,m})^{d+1}$$

for $\forall p$, $\forall d\geq 1$, and $0\leq l\leq m-1$.

Proof:

$$\begin{aligned}
&Q(C_t|Z_{t-(d+1)m-l}^p)\\
=&\sum_j Q(C_t|Z_{t-(d+1)m-l}^p, X_{t-(d+2)m}=j)Q(X_{t-(d+2)m}=j|Z_{t-(d+1)m-l}^p)
\end{aligned}$$

and therefore,

$$M_{d+2}^- \leq Q(C_t|Z_{t-(d+1)m-l}^p) \leq M_{d+2}^+$$

On the other hand,

$$\begin{aligned}
&Q(C_t|Z_{t-dm-l}^p)\\
=&\sum_{z_{t-(d+1)m-l}^{t-dm-l-1}} Q(C_t|Z_{t-(d+1)m-l}^p)Q(Z_{t-(d+1)m-l}^{t-dm-l-1}=z_{t-(d+1)m-l}^{t-dm-l-1}|Z_{t-dm-l}^p)
\end{aligned}$$

and thus,

$$M^-_{d+2} \leq Q(C_t|Z^p_{t-dm-l}) \leq M^+_{d+2}$$

Therefore, from Lemma 7, we have

$$|Q(C_t|Z^p_{t-dm-l}) - Q(C_t|Z^p_{t-(d+1)m-l})| \leq M^+_{d+2} - M^-_{d+2} \leq (\rho_{\delta,k,m})^{d+1}$$

Note that the result does not depend on either Q or l.

Appendix 2

Before proving Lemma 3 we need following lemma first. Part (b),(c), and (d) are crucial for Lemma 3, and Part (a) enables Part(b). Part (a) is the reason why we need a randomization of the filter.

Lemma 9 Suppose $Q \in \Theta_k^\delta$ and fix $\delta > 0$.

(a) We have

$$\|\hat{X}^\epsilon_Q(z^0_{-t_1}) - \hat{X}^\epsilon_Q(z^0_{-t_2})\|_1 \leq M^2 \cdot \|\mathbf{Q}_{X_0|z^0_{-t_1}} - \mathbf{Q}_{X_0|z^0_{-t_2}}\|_1,$$

where $t_1, t_2 > 0$ are arbitrary integers. That is, for any integer $t > 0$ and any individual sequence z^0_{-t}, $\hat{X}^\epsilon_Q(z^0_{-t})$ is a Lipschitz continuous function in $\mathbf{Q}_{X_0|z^0_{-t}}$.

(b) $\ell(X_0, \hat{X}^\epsilon_Q(Z^0_{-t})) \to \ell(X_0, \hat{X}^\epsilon_Q(Z^0_{-\infty}))$ *a.s.* uniformly on Θ_k^δ

(c) For $\forall Q \in \Theta_k^\delta$, and $\forall \omega$, $\exists \quad 0 < \gamma < 1, \beta > 0$, such that $|Q(X_0|Z^0_{-t}) - Q(X_0|Z^0_{-\infty})| < \beta\gamma^t$.

(d) For fixed t,$\eta > 0$, $\exists$ some finite set $\mathcal{F}_k(t,\eta) \subset \Theta_k^\delta$, such that

$$\max_{Q \in \Theta_k^\delta} \min_{Q' \in \mathcal{F}_k(t,\eta)} \max_{x_0, z^0_{-t}} |Q(x_0|z^0_{-t}) - Q'(x_0|z^0_{-t})| \leq \eta$$

Proof:

(a) For given simplex vector $\mathbf{Q}$, fixed $\hat{x}$, and B_ϵ defined as in Section 4-A, we define followings.

- $S_{\hat{x}}(\mathbf{Q}) \triangleq \{\mathbf{W} \in B_\epsilon : B(\mathbf{Q} + \mathbf{W}) = \hat{x}\}$
- $DP(\hat{x}) \triangleq \left\{\mathbf{c}^T\mathbf{y} = 0 : \mathbf{y} \in \mathbb{R}^M, \mathbf{c} = \lambda_{\hat{x}} - \lambda_a, \forall a \in \mathcal{A}\backslash\{\hat{x}\}\right\}$
- $\text{dist}(\mathbf{Q}, \mathbf{c}^T\mathbf{y} = 0) \triangleq$ The shortest L_2 distance from a simplex vector $\mathbf{Q}$ to the plane $\mathbf{c}^T\mathbf{y} = 0$

That is, $S_{\hat{x}}(\mathbf{Q})$ is a set of vectors in ϵ-ball, B_ϵ, that makes the Bayes response $B(\mathbf{Q} + \mathbf{W})$ equal to $\hat{x}$. Also, $DP(\hat{x})$ is a set of decision planes that separate the decision region for the reconstruction alphabet $\hat{x}$ and other alphabets. Then, for some fixed t, by definition,

$$\hat{X}^\epsilon_Q(z^0_{-t})[\hat{x}] = \frac{\text{Vol}(S_{\hat{x}}(\mathbf{Q}_{X_0|z^0_{-t}}))}{\text{Vol}(B_\epsilon)},$$

where Vol$(\cdot)$ is a volume of a set. Since Vol(B_ϵ) is a constant, for any t_1 and t_2, we have

$$|\hat{X}^\epsilon_Q(z^0_{-t_1})[\hat{x}] - \hat{X}^\epsilon_Q(z^0_{-t_2})[\hat{x}]| = \frac{|\text{Vol}(S_{\hat{x}}(\mathbf{Q}_{X_0|z^0_{-t_1}})) - \text{Vol}(S_{\hat{x}}(\mathbf{Q}_{X_0|z^0_{-t_2}}))|}{\text{Vol}(B_\epsilon)}. \quad (39)$$

For the numerater, as a crude bound, we get

$$\begin{aligned}
&|\mathrm{Vol}(S_{\hat{x}}(\mathbf{Q}_{X_0|z^0_{-t_1}})) - \mathrm{Vol}(S_{\hat{x}}(\mathbf{Q}_{X_0|z^0_{-t_2}}))| \\
\leq &\mathrm{Vol}(B_\epsilon^{M-1}) \cdot \sum_{\mathbf{c}^T\mathbf{y}=0 \in DP(\hat{x})} \Big| \mathrm{dist}(\mathbf{Q}_{X_0|z^0_{-t_1}}, \mathbf{c}^T\mathbf{y}=0) - \mathrm{dist}(\mathbf{Q}_{X_0|z^0_{-t_2}}, \mathbf{c}^T\mathbf{y}=0) \Big|,
\end{aligned} \tag{40}$$

where $B_\epsilon^{M-1} = \{\mathbf{U} \in \mathbb{R}^{M-1} : \|\mathbf{U}\|_2 \leq \epsilon\}$. Since

$$\mathrm{dist}(\mathbf{Q}, \mathbf{c}^T\mathbf{y}=0) = \frac{|\mathbf{c}^T\mathbf{Q}|}{\|\mathbf{c}\|_2},$$

we have

$$\begin{aligned}
&\mathrm{dist}(\mathbf{Q}_{X_0|z^0_{-t_1}}, \mathbf{c}^T\mathbf{y}=0) - \mathrm{dist}(\mathbf{Q}_{X_0|z^0_{-t_2}}, \mathbf{c}^T\mathbf{y}=0) \\
=&\frac{|\mathbf{c}^T\mathbf{Q}_{X_0|z^0_{-t_1}}| - |\mathbf{c}^T\mathbf{Q}_{X_0|z^0_{-t_2}}|}{\|\mathbf{c}\|_2} \\
\leq&\frac{\Big|\mathbf{c}^T(\mathbf{Q}_{X_0|z^0_{-t_1}} - \mathbf{Q}_{X_0|z^0_{-t_2}})\Big|}{\|\mathbf{c}\|_2} && (41) \\
\leq&\|\mathbf{Q}_{X_0|z^0_{-t_1}} - \mathbf{Q}_{X_0|z^0_{-t_2}}\|_2 && (42) \\
\leq&\|\mathbf{Q}_{X_0|z^0_{-t_1}} - \mathbf{Q}_{X_0|z^0_{-t_2}}\|_1 && (43)
\end{aligned}$$

where (41) is from the triangular inequality, (42) is from Cauchy-Schwartz inequality, and (43) is from the fact that L_2-norm is less than or equal to L_1-norm. Therefore, (40) becomes

$$|\mathrm{Vol}(S_{\hat{x}}(\mathbf{Q}_{X_0|z^0_{-t_1}}) - \mathrm{Vol}(S_{\hat{x}}(\mathbf{Q}_{X_0|z^0_{-t_2}})| \leq M \cdot \mathrm{Vol}(B_\epsilon^{M-1}) \cdot \|\mathbf{Q}_{X_0|z^0_{-t_1}} - \mathbf{Q}_{X_0|z^0_{-t_2}}\|_1,$$

and thus, (39) becomes

$$\begin{aligned}
|\hat{X}_Q^\epsilon(z^0_{-t_1})[\hat{x}] - \hat{X}_Q^\epsilon(z^0_{-t_2})[\hat{x}]| \leq &M \cdot \frac{\mathrm{Vol}(B_\epsilon^{M-1})}{\mathrm{Vol}(B_\epsilon)} \cdot \|\mathbf{Q}(X_0|z^0_{-t_1}) - \mathbf{Q}(X_0|z^0_{-t_2})\|_1. \\
\leq &M \cdot \|\mathbf{Q}_{X_0|z^0_{-t_1}} - \mathbf{Q}_{X_0|z^0_{-t_2}}\|_1.
\end{aligned}$$

Therefore, we have

$$\|\hat{X}_Q^\epsilon(z^0_{-t_1}) - \hat{X}_Q^\epsilon(z^0_{-t_2})\|_1 \leq M^2 \cdot \|\mathbf{Q}_{X_0|z^0_{-t_1}} - \mathbf{Q}_{X_0|z^0_{-t_2}}\|_1,$$

and Part (a) is proved.

(b) By the exact same argument as in proving Lemma 1, we can easily know that $Q(X_0|Z^0_{-t}) \to Q(X_0|Z^0_{-\infty})$ for $\forall\omega$, uniformly in $\forall Q \in \Theta_k^{\delta_k}$. Since we have

$$\begin{aligned}
&\Big|\ell(X_0, \hat{X}_Q^\epsilon(Z^0_{-t})) - \ell(X_0, \hat{X}_Q^\epsilon(Z^0_{-\infty}))\Big| \\
=&\Big|\sum_{\hat{x}} \Lambda(X_0, \hat{x})\Big(\hat{X}_Q^\epsilon(Z^0_{-t})[\hat{x}] - \hat{X}_Q^\epsilon(Z^0_{-\infty})[\hat{x}]\Big)\Big| \\
\leq&\Lambda_{\max}\|\hat{X}_Q^\epsilon(Z^0_{-t}) - \hat{X}_Q^\epsilon(Z^0_{-\infty})\|_1 \\
\leq&\Lambda_{\max}M^2 \cdot \|\mathbf{Q}(X_0|Z^0_{-t}) - \mathbf{Q}(X_0|Z^0_{-\infty})\|_1,
\end{aligned}$$

we get the uniform convergence.

(c) Again, let's follow the argument in the proof of Lemma 1. Suppose $t = jk + l$, where $j = \lfloor t/k \rfloor$, and $l = t \mod k$. Then,

$$
\begin{aligned}
&|Q(X_0|Z_{-t}^0) - Q(X_0|Z_{-\infty}^0)| \\
=&|Q(X_0|Z_{-jk-l}^0) - Q(X_0|Z_{-\infty}^0)| \\
\leq&\sum_{i=j}^{\infty} |Q(X_0|Z_{-ik-l}^0) - Q(X_0|Z_{-(i+1)k-l}^0)| \\
\leq&\sum_{i=j}^{\infty} \rho^{i+1} && (44)\\
=&\frac{\rho^{j+1}}{1-\rho} = \frac{\rho}{1-\rho}\rho^{\lfloor t/k \rfloor} = \frac{\rho^{1-\frac{l}{k}}}{1-\rho}(\rho^{1/k})^t && (45)\\
\leq&\frac{1}{1-\rho}(\rho^{1/k})^t && (46)
\end{aligned}
$$

where $\rho = \rho_{\delta,k,k}$ as defined in Lemma 7, and (44) follows from Lemma 8. By letting $\beta = \frac{1}{1-\rho}$, and $\gamma = \rho^{1/k}$, we have proved Part (c).

(d) We know that for the individual sequence pair (x_0, z_{-t}^0),

$$
\begin{aligned}
Q(x_0|z_{-t}^0) =& \frac{\sum_{x_{-t}^{-1}} Q(x_{-t}^0, z_{-t}^0)}{Q(z_{-t}^0)} \\
=& \frac{\sum_{x_{-t}^{-1}} Q(x_{-t}^0, z_{-t}^0)}{\sum_{x_{-t}^0} Q(x_{-t}^0, z_{-t}^0)} \\
=& \frac{\sum_{x_{-t}^{-1}} Q(x_{-t}^0)Q(z_{-t}^0|x_{-t}^0)}{\sum_{x_{-t}^0} Q(x_{-t}^0)Q(z_{-t}^0|x_{-t}^0)} \\
=& \frac{\sum_{x_{-t}^{-1}} \left(Q(x_{-t}^0)\prod_{i=-t}^0 \Pi(x_i,z_i)\right)}{\sum_{x_{-t}^0} \left(Q(x_{-t}^0)\prod_{i=-t}^0 \Pi(x_i,z_i)\right)}.
\end{aligned}
$$

For $Q \in \Theta_k^\delta$, $\mathbf{\Pi}$ is fixed and we can think of $\prod_{i=-t}^0 \Pi(x_i, z_i)$ as a constant for the individual sequence pair (x_{-t}^0, z_{-t}^0). Since

$$
Q(x_{-t}^0) = Q(x_{-t}^{k-1-t}) \prod_{j=k-t}^{0} a_{x_{j-k}^{j-1} x_{j-k+1}^{j}},
$$

$Q(x_0|z_{-t}^0)$ is the ratio of two finite order polynomials of $\{a_{ij}\}$, and as Θ_k^δ is closed and bounded, $Q(x_0|z_{-t}^0)$ is a uniformly continuous function of $\{a_{ij}\}$. Therefore, for given η, $\exists \epsilon(\eta)$ such that $\|Q - Q^{'}\|_1 < \epsilon(\eta)$ implies

$$
\max_{x_0, z_{-t}^0} |Q(x_0|z_{-t}^0) - Q'(x_0|z_{-t}^0)| \leq \eta,
$$

since there are only finite number of possible (x_0, z_{-t}^0) pairs. Also, since Θ_k^δ is compact, we can always find a finite set, $\mathcal{F}_k(t,\eta)$ that for any $Q \in \Theta_k^\delta$, there exists at least one $Q' \in \mathcal{F}_k(t,\eta)$, that satisfies $\|Q - Q^{'}\|_1 < \epsilon(\eta)$. Therefore, Part (d) is proved.

Proof of Lemma 3: To prove Lemma 3, first consider following limit.

$$\begin{aligned}
&\lim_{n\to\infty} E\Big(L_{\hat{\mathbf{X}}^{\epsilon}_Q}(X^n, Z^n)\Big) \\
=&\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} E\Big(\ell(X_t, \hat{X}^{\epsilon}_Q(Z^t))\Big) \\
=&\lim_{t\to\infty} E\Big(\ell(X_t, \hat{X}^{\epsilon}_Q(Z^t))\Big) && (47)\\
=&\lim_{t\to\infty} E\Big(\ell(X_0, \hat{X}^{\epsilon}_Q(Z^0_{-(t-1)}))\Big) && (48)\\
=&E\Big(\ell(X_0, \hat{X}^{\epsilon}_Q(Z^0_{-\infty}))\Big) \text{ uniformly on } \Theta^{\delta}_k, && (49)
\end{aligned}$$

where (47) is from Cesáro's mean convergence theorem, (48) is from stationarity, and (49) is from Lemma 9(b) and bounded convergence theorem. Thus, to complete the proof, we need to show that

$$\lim_{n\to\infty} L_{\hat{\mathbf{X}}^{\epsilon}_Q}(X^n, Z^n) = E\Big(\ell(X_0, \hat{X}^{\epsilon}_Q(Z^0_{-\infty}))\Big) \quad a.s. \quad \text{uniformly on } \Theta^{\delta}_k \tag{50}$$

Now, let's show the pointwise convergence in (50) without the uniformity by using ergodic theorem. For given Q, define

$$\begin{aligned}
g_{t,Q}(X,Z) &\triangleq \ell(X_0, \hat{X}^{\epsilon}_Q(Z^0_{-(t-1)})) \\
g_Q(X,Z) &\triangleq \ell(X_0, \hat{X}^{\epsilon}_Q(Z^0_{-\infty}))
\end{aligned}$$

and denote by T the shift operator. Then, what we should prove becomes

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} g_{t,Q}(T^t(X,Z)) = E\Big(g_Q(X,Z)\Big) \quad a.s.$$

while the ergodic theorem gives

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} g_{Q}(T^t(X,Z)) = E\Big(g_Q(X,Z)\Big) \quad a.s.$$

Observe that

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{t=1}^{n} g_{t,Q}(T^t(X,Z)) - \frac{1}{n}\sum_{t=1}^{n} g_Q(T^t(X,Z))\Big| \\
\le&\frac{1}{n}\sum_{t=1}^{n}\Big|g_{t,Q}(T^t(X,Z)) - g_Q(T^t(X,Z))\Big| \\
=&\frac{1}{n}\sum_{t=1}^{n}\Big|\ell(X_t, \hat{X}^{\epsilon}_Q(Z^t_1)) - \ell(X_t, \hat{X}^{\epsilon}_Q(Z^t_{-\infty}))\Big|.
\end{aligned}$$

Since Lemma 9(c) holds for $\forall\omega$, we can think that the lemma holds for all individual sequence pair $(x_0, z^0_{-\infty})$. Thus, it holds for all individual pair $(x_t, z^t_{-\infty})$, too, and we can conclude that $Q(X_t|Z^t_1) \to Q(X_t|Z^t_{-\infty})$ for $\forall\omega$ as $t\to\infty$. Hence, by exactly the same argument as Lemma 9(a) and Lemma 9(b), we conclude that $\ell(X_t, \hat{X}^{\epsilon}_Q(Z^t_1)) \to \ell(X_t, \hat{X}^{\epsilon}_Q(Z^t_{-\infty}))$ almost surely as $t\to\infty$. Now, by Cesáro's mean convergence theorem , we obtain

$$\frac{1}{n}\sum_{t=1}^{n}\Big|\ell(X_t, \hat{X}^{\epsilon}_Q(Z^t_1)) - \ell(X_t, \hat{X}^{\epsilon}_Q(Z^t_{-\infty}))\Big| \to 0 \quad a.s.$$

Therefore, we get

$$L_{\hat{\mathbf{X}}^{\epsilon}_{Q}}(X^n, Z^n) \rightarrow E\Big(\ell(X_0, \hat{X}^{\epsilon}_{Q}(Z^0_{-\infty}))\Big) \quad a.s.$$

Note that up to this point we cannot guarantee the uniformity of the convergence, since the ergodic theorem only gives the individual convergence for each Q. To show the uniformity of the convergence in (50), first define the following quantity for some fixed integer $t \in [1, n-1]$,

$$L_{\hat{\mathbf{X}}^{\epsilon}_{Q,t}}(X^n, Z^n) = \frac{1}{n}\Big(\sum_{i=1}^{t} \ell(X_i, \hat{X}^{\epsilon}_{Q}(Z^i)) + \sum_{i=t+1}^{n} \ell(X_i, \hat{X}^{\epsilon}_{Q}(Z^i_{i-t}))\Big).$$

From Lemma 9(d), for any $Q \in \Theta^{\delta}_{k}$ and fixed $t, \eta > 0$, we can pick some $Q' \in \mathcal{F}_k(t,\eta)$ such that $\|Q - Q^{'}\|_1 < \epsilon(\eta)$, and thus,

$$\max_{x_0, z^0_{-t}} |Q(x_0|z^0_{-t}) - Q'(x_0|z^0_{-t})| \leq \eta.$$

By adding and subtracting some common terms involving such $Q^{'}$, and from the triangle inequality, we have,

$$\begin{aligned}
&\Big|L_{\hat{\mathbf{X}}^{\epsilon}_{Q}}(X^n, Z^n) - E\Big(\ell(X_0, \hat{X}^{\epsilon}_{Q}(Z^0_{-\infty}))\Big)\Big| \\
\leq &\Big|L_{\hat{\mathbf{X}}^{\epsilon}_{Q}}(X^n, Z^n) - L_{\hat{\mathbf{X}}^{\epsilon}_{Q,t}}(X^n, Z^n)\Big| + \Big|L_{\hat{\mathbf{X}}^{\epsilon}_{Q,t}}(X^n, Z^n) - L_{\hat{\mathbf{X}}^{\epsilon}_{Q',t}}(X^n, Z^n)\Big| + \Big|L_{\hat{\mathbf{X}}^{\epsilon}_{Q',t}}(X^n, Z^n) - L_{\hat{\mathbf{X}}^{\epsilon}_{Q'}}(X^n, Z^n)\Big| \\
+ &\Big|L_{\hat{\mathbf{X}}^{\epsilon}_{Q'}}(X^n, Z^n) - E\Big(\ell(X_0, \hat{X}^{\epsilon}_{Q'}(Z^0_{-\infty}))\Big)\Big| + \Big|E\Big(\ell(X_0, \hat{X}^{\epsilon}_{Q'}(Z^0_{-\infty}))\Big) - E\Big(\ell(X_0, \hat{X}^{\epsilon}_{Q}(Z^0_{-\infty}))\Big)\Big| \qquad (51)
\end{aligned}$$

Now, the goal becomes to show that the terms in the righthand side of the inequality converges to zero independent of Q as n, t, and η varies. First, we will bound each term, and send $n \rightarrow \infty$.

(1)

$$\begin{aligned}
&\Big|L_{\hat{\mathbf{X}}^{\epsilon}_{Q}}(X^n, Z^n) - L_{\hat{\mathbf{X}}^{\epsilon}_{Q,t}}(X^n, Z^n)\Big| \\
\leq &\frac{1}{n}\sum_{i=t+1}^{n}\Big|\ell(X_i, \hat{X}^{\epsilon}_{Q}(Z^i)) - \ell(X_i, \hat{X}^{\epsilon}_{Q}(Z^i_{i-t}))\Big| \\
\leq &\Lambda_{\max} \cdot \frac{1}{n}\sum_{i=t+1}^{n} \|\hat{X}^{\epsilon}_{Q}(Z^i) - \hat{X}^{\epsilon}_{Q}(Z^i_{i-t})\|_1 \\
\leq &\Lambda_{\max} M^2 \cdot \frac{1}{n}\sum_{i=t+1}^{n} \|\mathbf{Q}_{X_0|Z^0_{-i}} - \mathbf{Q}_{X_0|Z^0_{-t}}\|_1 \qquad (52) \\
\leq &\Lambda_{\max} M^3 \cdot \frac{1}{n}\sum_{i=t+1}^{n} (\beta\gamma^t + \beta\gamma^i) \qquad (53) \\
\rightarrow &\Lambda_{\max} M^3 \beta\gamma^t \quad a.s. \text{ uniformly on } \Theta^{\delta}_{k} \qquad (54)
\end{aligned}$$

where (52) is from stationarity and Lemma 9(a), (53) is from Lemma 9(c), and (54) is from the Cesáro's mean convergence theorem. Since (53) does not depend on Q, the limit is uniform on Θ^{δ}_{k}.

(2)

$$
\begin{aligned}
&\Big|L_{\hat{\mathbf{X}}^{\epsilon}_{Q,t}}(X^n,Z^n)-L_{\hat{\mathbf{X}}^{\epsilon}_{Q',t}}(X^n,Z^n)\Big|\\
\leq&\frac{1}{n}\sum_{i=t+1}^{n}|\ell(X_i,\hat{X}^{\epsilon}_{Q}(Z_{i-t}^{i}))-\ell(X_i,\hat{X}^{\epsilon}_{Q'}(Z_{i-t}^{i}))|+\frac{t}{n}\cdot\Lambda_{max}\\
\leq&\Lambda_{\max}\cdot\frac{1}{n}\sum_{i=t+1}^{n}\|\hat{X}^{\epsilon}_{Q}(Z_{i-t}^{i})-\hat{X}^{\epsilon}_{Q'}(Z_{i-t}^{i})\|_1+\frac{t}{n}\cdot\Lambda_{max}\\
\leq&\Lambda_{\max}M^2\cdot\frac{1}{n}\sum_{i=t+1}^{n}\|\mathbf{Q}_{X_i|Z_{i-t}^{i}}-\mathbf{Q'}_{X_i|Z_{i-t}^{i}}\|_1+\frac{t}{n}\cdot\Lambda_{max} \qquad (55)\\
\leq&\Lambda_{\max}M^3\frac{n-t}{n}\cdot\eta+\frac{t}{n}\cdot\Lambda_{max} \qquad (56)\\
\to&\Lambda_{\max}M^3\eta \quad a.s. \text{ uniformly on } \Theta_k^{\delta}
\end{aligned}
$$

where (55) is from Lemma 9(a), and (56) is from Lemma 9(d). Since (56) does not depend on Q, the limit is also uniform on Θ_k^{δ}.

(3)

$$
\Big|L_{\hat{\mathbf{X}}^{\epsilon}_{Q',t}}(X^n,Z^n)-L_{\hat{\mathbf{X}}^{\epsilon}_{Q'}}(X^n,Z^n)\Big|\to\Lambda_{max}M^3\beta\gamma^t \quad a.s.
$$

by following the same argument as (1). Since $\mathcal{F}_k(t,\eta)$ is finite, this convergence is uniform on $\mathcal{F}_k(t,\eta)$.

(4)

$$
\Big|L_{\hat{\mathbf{X}}^{\epsilon}_{Q'}}(X^n,Z^n)-E\Big(\ell(X_0,\hat{X}^{\epsilon}_{Q'}(Z_{-\infty}^{0}))\Big)\Big|\to 0 \quad a.s.
$$

from the proof of pointwise convergence above. As in (3), this convergence is also uniform on $\mathcal{F}_k(t,\eta)$.

(5)

$$
\begin{aligned}
&\Big|E\Big(\ell(X_0,\hat{X}^{\epsilon}_{Q'}(Z_{-\infty}^{0}))\Big)-E\Big(\ell(X_0,\hat{X}^{\epsilon}_{Q}(Z_{-\infty}^{0}))\Big)\Big|\\
\leq&\Big|E\Big(\ell(X_0,\hat{X}^{\epsilon}_{Q'}(Z_{-\infty}^{0}))\Big)-E\Big(\ell(X_0,\hat{X}^{\epsilon}_{Q'}(Z_{-t}^{0}))\Big)\Big|+\Big|E\Big(\ell(X_0,\hat{X}^{\epsilon}_{Q'}(Z_{-t}^{0}))\Big)-E[\ell(X_0,\hat{X}^{\epsilon}_{Q}(Z_{-t}^{0}))\Big)\Big|\\
+&\Big|E\Big(\ell(X_0,\hat{X}^{\epsilon}_{Q}(Z_{-t}^{0}))\Big)-E\Big(\ell(X_0,\hat{X}^{\epsilon}_{Q}(Z_{-\infty}^{0}))\Big)\Big|\\
\leq&\sum_{x_0,z_{-\infty}^{0}}P(x_0,z_{-\infty}^{0})\Big|\ell(x_0,\hat{X}^{\epsilon}_{Q'}(z_{-\infty}^{0}))-\ell(x_0,\hat{X}^{\epsilon}_{Q'}(z_{-t}^{0}))\Big|+\sum_{x_0,z_{-t}^{0}}P(x_0,z_{-t}^{0})\Big|\ell(x_0,\hat{X}^{\epsilon}_{Q'}(z_{-t}^{0}))-\ell(x_0,\hat{X}^{\epsilon}_{Q}(z_{-t}^{0}))\Big|\\
+&\sum_{x_0,z_{-\infty}^{0}}P(x_0,z_{-\infty}^{0})\Big|\ell(x_0,\hat{X}^{\epsilon}_{Q}(z_{-\infty}^{0}))-\ell(x_0,\hat{X}^{\epsilon}_{Q}(z_{-t}^{0}))\Big|\\
\leq&\Lambda_{max}M^3\Big(2\beta\gamma^t+\eta\Big),
\end{aligned}
$$

by similar argument as in (1) and (2).

Therefore, by taking limit supremum on both side of (51), we get

$$
\begin{aligned}
&\limsup_{n\to\infty}\Big|L_{\hat{\mathbf{X}}^{\epsilon}_{Q}}(X^n,Z^n)-E\Big(\ell(X_0,\hat{X}^{\epsilon}_{Q}(Z_{-\infty}^{0}))\Big)\Big|\\
\leq&\Lambda_{max}M^3\Big(4\beta\gamma^t+2\eta\Big) \quad a.s. \quad \text{uniformly on } \Theta_k^{\delta}.
\end{aligned}
$$

Since t and η are arbitrary, by sending $t \to \infty$ and $\eta \downarrow 0$, we have

$$\limsup_{n\to\infty} \left| L_{\hat{\mathbf{X}}^{\epsilon}_{Q}}(X^n, Z^n) - E\Big(\ell(X_0, \hat{X}^{\epsilon}_{Q}(Z^0_{-\infty}))\Big)\right| \leq 0 \quad a.s. \quad \text{uniformly on } \Theta^{\delta}_{k}.$$

Therefore, the lemma is proved. ■

Appendix 3

Here, we prove Corollary 1.

Proof of Corollary 1: First note the subtle point that Corollary 1 does not directly follow from Lemma 3. Since the probability law Q^t_k that we are using to filter each block is changing every block, whereas the uniform convergence in Lemma 3 is for the fixed $Q \in \Theta^{\delta_k}_{k}$ for all t, it is not enough to guarantee the Corollary. However, since Q^t_k remains the same within each block, we can still use the result of Lemma 3 if the block length gets long enough. Keeping this in mind, let's take a more careful look at each block. In the proof, for the brevity of notation, let's denote

$$\ell_t(Q) \triangleq \ell(X_t, \hat{X}^{\epsilon}_{Q}(Z^t)),$$

since we are always dealing with the randomized filter, and there is no possibility of confusion. Now, fix any $\delta > 0$. Then, from (4),

$$\exists I, \quad \text{such that} \quad \frac{m_{I-1}}{m_I} < \frac{\delta}{8\ell_{max}},$$

and from Lemma 3,

$$\exists N, \quad \text{such that} \quad \max_{Q\in\Theta^{\delta_k}_{K}} \left| L_{\hat{\mathbf{X}}^{\epsilon}_{Q}}(X^n, Z^n) - EL_{\hat{\mathbf{X}}^{\epsilon}_{Q}}(X^n, Z^n)\right| < \delta/4.$$

Recalling the definition $i(t) \triangleq \max\{i : m_i \leq t\}$, we let $I_0 = \max(I, i(N)+1)$. Then, for any $n \geq m_{I_0}$, and $m_{i(n)} \leq n < m_{i(n)+1}$,

$$\left| L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n, Z^n) - \hat{E}L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n, Z^n)\right| \tag{57}$$

$$\leq \frac{1}{n}\left|\sum_{t=1}^{m_{i(n)-1}} \Big(\ell_t(Q^t_k) - \hat{E}(\ell_t(Q^t_k))\Big)\right| + \frac{1}{n}\left|\sum_{t=m_{i(n)-1}+1}^{m_{i(n)}} \Big(\ell_t(\hat{Q}[Z^{m_{i(n)-1}}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)-1}}]))\Big)\right| \tag{58}$$

$$+\frac{1}{n}\left|\sum_{t=m_{i(n)}+1}^{n} \Big(\ell_t(\hat{Q}[Z^{m_{i(n)}}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)}}]))\Big)\right|. \tag{59}$$

Note that in the second and third term, Q^t_k is fixed to $\hat{Q}[Z^{m_{i(n)-1}}]$ and $\hat{Q}[Z^{m_{i(n)}}]$ from the definition of our filter. Now, we can bound each term. For the first term, since $n \geq m_{i(n)} \geq m_I$, we know that $\frac{m_{i(n)-1}}{n} \leq \frac{m_{i(n)-1}}{m_{i(n)}} < \frac{\delta}{8\ell_{max}}$. Therefore,

$$\frac{1}{n}\left|\sum_{t=1}^{m_{i(n)-1}} \Big(\ell_t(Q^t_k) - \hat{E}(\ell_t(Q^t_k))\Big)\right| \leq \frac{\delta}{8\ell_{max}} \cdot \ell_{max} = \frac{\delta}{8}.$$

For the second term, since $n \geq m_{i(n)} \geq N$,

$$\frac{1}{n}\left|\sum_{t=m_{i(n)-1}+1}^{m_{i(n)}} \Big(\ell_t(\hat{Q}[Z^{m_{i(n)-1}}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)-1}}]))\Big)\right| \tag{60}$$

$$\leq \frac{m_{i(n)}}{n}\frac{1}{m_{i(n)}}\left|\sum_{t=1}^{m_{i(n)}} \Big(\ell_t(\hat{Q}[Z^{m_{i(n)-1}}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)-1}}]))\Big)\right| + \frac{1}{n}\left|\sum_{t=1}^{m_{i(n)-1}} \Big(\ell_t(\hat{Q}[Z^{m_{i(n)-1}}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)-1}}]))\Big)\right| \tag{61}$$

$$\leq \frac{\delta}{4} + \frac{\delta}{8\ell_{max}} \cdot \ell_{max} = \frac{3\delta}{8} \tag{62}$$

Finally, for the last term,

$$\frac{1}{n}\left|\sum_{t=m_{i(n)}+1}^{n} \Big(\ell_t(\hat{Q}[Z^{m_{i(n)}}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)}}]))\Big)\right| \tag{63}$$

$$\leq \frac{1}{n}\left|\sum_{t=1}^{n} \Big(\ell_t(\hat{Q}[Z^{m_{i(n)}}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)}}]))\Big)\right| + \frac{1}{n}\left|\sum_{t=1}^{m_{i(n)}} \Big(\ell_t(\hat{Q}[Z^{m_{i(n)}}]) - \hat{E}(\ell_t(\hat{Q}[Z^{m_{i(n)}}]))\Big)\right| \tag{64}$$

$$\leq \frac{\delta}{4} + \frac{\delta}{4} = \frac{\delta}{2}. \tag{65}$$

Therefore, for any $n \geq m_{I_0}$, and $m_{i(n)} \leq n \leq m_{i(n)+1}$, we have

$$\left| L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n, Z^n) - \hat{E}L_{\hat{\mathbf{X}}^{\epsilon}_{univ,k}}(X^n, Z^n)\right| < \delta,$$

and since δ was arbitrary, we have the corollary. ■

References

[1] L.E. Baum and T. Petrie, "Statistical Inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol.37, 1554-1563, 1966

[2] L.E. Baum, T.Petrie, G. Soules, and N. Weiss, "A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, 41:164-171, 1970

[3] P.J. Bickel, Y. Ritov, and T. Rydén, "Asymptotic normality of the maxumum-likelihood estimator for general hidden Markov models," *Ann. Statist.*, 26(4):1614-1635, 1998

[4] R.W. Chang and J.C. Hancock, "On receiver structures for channels having memory," *IEEE Trans. Inform. Theory*, vol.IT-12:463-468, Oct 1996

[5] T.M. Cover and J.A. Thomas, "Elements of Information Theory," *New York: Wiley, 1991*

[6] A. Dembo and T. Weissman, "Universal denoising for the finite-input-general-output channel," *IEEE Trans. Inform. Theory*, 51(4):1507-1517, April 2005

[7] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inform. Theory*, 48(6):1518-1569, June 2002

[8] L. Finesso, "Consistent Estimation of the Order for Markov and Hidden Markov Chains," *Ph.D Dissertation, Univ. Maryland, College Park*, 1990

[9] G. Gemelos, S. Sigurjónsson, and T. Weissman, "Universal discrete denoising under channel uncertainty," *Proceedings of Int. Symp. Inf. Th.* , pp.199, Chicago, IL, Jun./Jul. 2004

[10] G.H. Golub and C.D. Meyer, "Using the QR factorization and group inversion to compute, differentiate, and estimate the sensitivity of stationary probabilities for Markov chains" *SIAM J. Algebraic Discrete Meth.*, 7, pp.273-281, 1986

[11] J.C. Kieffer, "Strongly consistent code-based identification and order estimation for constrained finite-state model classes," *IEEE Trans. Inform. Theory*, 39:893-902, May 1993

[12] B.G. Leroux, "Maximum-likelihood estimation for hidden Markov models," *Stochastic Processes Their Appic.*, vol. 40, pp.127-143, 1992

[13] C.-C Liu and P. Narayan, "Order estimation and sequential universal data compression of a hidden Markov source bu the model of mixtures," *IEEE Trans. Inform. Theory*,40:1167-1180, July 1994

[14] N. Merhav and M. Feder, "Universal Prediction," *IEEE Trans. Inform. Theory*, 44(6):2124-2147, Oct 1998

[15] N. Feder, N. Merhav, and M. Gutman, "Universal Prediction for individual sequences" *IEEE Trans. Inform. Theory*, 38:1258-1270, July 1992

[16] L. Mevel and L. Finesso, "Asymptotical statistics of misspecified hidden Markov models," *IEEE Trans. Automatic Control*, 49(7):1123-1132, Jul 2004

[17] T. Moon and T. Weissman, "Discrete universal filtering via hidden Markov modelling," *Int. Symp. Inf. Th.* , Adelaide, Australia, September 2005.

[18] E. Ordentlich, T. Weissman, M. Weinberger, A. Somekh-Baruch and N. Merhav, "Discrete universal filtering through incremental parsing" *Data Compression Conference (DCC 2004)*, p. 352-361, Snowbird, Utah March 23-25, 2004.

[19] E. Ordentlich, T. Weissman, M. Weinberger, A. Somekh-Baruch and N. Merhav, "Universal filtering via prediction", available at `http://www.stanford.edu/~tsachy/filtering.pdf`

[20] T. Rydén, "Estimating the order of hidden Markov models," *Statistics*, 26:345-354, 1995

[21] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. Inform. Theory*, 51(1):5-28, January 2005.

[22] R. Zhang and T. Weissman, "Discrete Denoising for Channels with Memory", *Communications in Informations and Systems*, 5(2):257-288, 2005

[23] J. Ziv and N. Merhav, "Estimating the number of states of a finite-state source," *IEEE Trans. Inform. Theory*,38:61-65, Jan 1992